UNIVERSITY OF JOHANNESBURG

COPYRIGHT AND CITATION CONSIDERATIONS FOR THIS THESIS/ DISSERTATION

- Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- NonCommercial — You may not use the material for commercial purposes.
- ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

How to cite this thesis

Surname, Initial(s). (2012). Title of the thesis or dissertation (Doctoral Thesis / Master's Dissertation). Johannesburg: University of Johannesburg. Available from: http://hdl.handle.net/102000/0002 (Accessed: 22 August 2017).

Improved Machine Learning Methods for Classification of Imbalanced Data

by

Sarah Alexandria Ebiaredoh-Mienye

A Dissertation submitted in partial fulfilment of the requirements for the degree

Master of Engineering

in

Electrical and Electronic Engineering

in the

Faculty of Engineering and the Built Environment

at the

University of Johannesburg

SUPERVISOR: Prof. Theo G. Swart

CO-SUPERVISOR: Dr. Ebenezer Esenogho

February 2021

Dedication

This dissertation is first of all dedicated to the Almighty God for the grace, strength and knowledge to complete it. And to my sponsor, Dr Nimibofa Ayawei, I am forever grateful for this opportunity and do not take it for granted. I also want to dedicate this work to my beloved husband, Ibomoiye Domor Mienye, for his support, encouragement, hours of proofreading, and most of all, for being my best cheerleader. Lastly, to my beautiful daughter for being a source of inspiration.

UNIVERSITY
OF
JOHANNESBURG

Declaration

I, **Sarah Alexandria Ebiaredoh-Mienye,** hereby declare that this dissertation is entirely my own work and has not been submitted anywhere else for academic credit either by myself or another person. I understand what plagiarism implies and declare that this dissertation is my own ideas, words, phrases, arguments, graphics, figures, results, and organisation except where reference is explicitly made to another's work. I understand further that any unethical academic behaviour, which includes plagiarism, is seen in a serious light by the University of Johannesburg and is punishable by disciplinary action.

Signed: Sarah Alexandria Ebiaredoh-Mienye Date: 22/02/2021

UNIVERSITY OF JOHANNESBURG

Acknowledgement

First of all, I would like to thank my supervisors, Prof Theo G. Swart and Dr Ebenezer Esenogho, for providing the needed support and guidance. I would also like to acknowledge my husband, Ibomoiye Domor Mienye, for your support throughout my research program. Finally, I must express my very profound gratitude to my daughter, family, and friends for providing me with constant support and continuous encouragement all through my research.

UNIVERSITY OF JOHANNESBURG

Abstract

The emergence of Big Data and machine learning (ML) has paved the way for numerous scientific advancements. A challenge which has hindered the progress and application of machine learning algorithms for certain classification tasks is the class imbalance problem. Imbalanced classification is a situation where there is a skewed distribution of the target variables. The class imbalance problem exists in several domains, including medical diagnosis, credit risk prediction, fraud detection, and other areas in which negatively labelled samples considerably exceeds the positively labelled samples. Using imbalanced data to train ML models often results in poor performance.

Several research works have proposed diverse methods to mitigate the class imbalance problem, including data sampling, ensemble learning, and feature learning. However, in this research, the focus is on effective feature learning. This dissertation presents two ML methods that are implemented to enhance the performance of diverse classifiers using publicly available imbalanced datasets.

- Firstly, a thorough literature review is conducted on various ML algorithms developed to solve the class imbalance problem.
- Secondly, a method was developed to improve the classification performance of some classifiers using stacked sparse autoencoder, with application to credit risk prediction.
- Thirdly, a method was introduced for medical diagnosis using an enhanced sparse autoencoder and softmax regression.

The methods implemented in this research outperformed most machine learning algorithms and scholarly works. Furthermore, this research work demonstrates the effect of effective feature learning on the performance of classifiers and the importance of training these classifiers with relevant data.

Keywords: Artificial neural network, autoencoder, deep learning, feature learning, machine learning, medical diagnosis

UNIVERSITY OF JOHANNESBURG

List of Abbreviations

Adam	Adaptive moment estimation algorithm
AE	Autoencoder
AI	Artificial intelligence
ANN	Artificial neural network
AUC	Area under the receiver operating characteristic curve
BPNN	Backpropagation neural network
CAD	Computer-aided diagnosis
CART	Classification and regression tree
CHD	Coronary heart disease
CKD	Chronic kidney disease
CNN	Convolutional neural network
CT	Computed tomography
CVD	Cardiovascular disease
DL	Deep learning
DNN	Deep neural network
DT	Decision tree
FPR	False positive rate
HD	Heart disease
HPV	Human papillomavirus
KL	Kullback-Leibler divergence
KNN	k-Nearest neighbor
LDA	Linear discriminant analysis
LR	Logistic regression
MDA	Marginalized autoencoder
ML	Machine learning
MSE	Mean squared error
NB	Naïve Bayes
PCA	Principal component analysis
PSD	Predictive sparse decomposition
PSO	Particle swarm optimization

UNIVERSITY OF JOHANNESBURG

ResNet	Residual neural network
RF	Random forest
RGB	Red, Green, Blue
ROC	Receiver operating characteristic curve
SAE	Sparse autoencoder
SGD	Stochastic gradient descent
SR	Softmax regression
SRAG	Stacked robust autoencoder with graph regularization
SSAE	Stacked sparse autoencoder
SSIM	Structural similarity index
SVM	Support vector machine
Tanh	Hyperbolic tangent activation function
TCIA	The Cancer Imaging Archive
TPR	True positive rate
UAMS	University of Arkansas for Medical Sciences
UCI	University of California, Irvine
WAE	Weighted aging classifier ensemble

UNIVERSITY OF JOHANNESBURG

Table of Contents

UNIVERSITY OF JOHANNESBURG

UNIVERSITY OF JOHANNESBURG

UNIVERSITY
OF
JOHANNESBURG

List of Figures

UNIVERSITY OF JOHANNESBURG

List of Tables

UNIVERSITY OF JOHANNESBURG

List of Publications

1. Ebiaredoh-Mienye, S. A., Esenogho, E. and Swart, T. G. (2020) 'Integrating Enhanced Sparse Autoencoder-Based Artificial Neural Network Technique and Softmax Regression for Medical Diagnosis', *Electronics*, 9(11), p. 1963. doi: 10.3390/electronics9111963. (As presented in Chapter 5)

2. Ebiaredoh-Mienye, S. A., Esenogho, E. and Swart, T. G., 'Artificial Neural Network Technique for Improving the Prediction of Credit Card Default: A Stacked Sparse Autoencoder Approach', *International Journal of Electrical and Computer Engineering (IJECE)*, Status: Accepted for publication. (As presented in Chapter 4)

UNIVERSITY OF JOHANNESBURG

CHAPTER 1

INTRODUCTION

1.1. Background

In recent years, there have been several advances in artificial intelligence (AI). These AI-based systems have outperformed humans in numerous applications such as medical diagnosis, speech recognition, image recognition, and gaming. AI is simply intelligence exhibited by machines, as opposed to the natural intelligence present in humans [1]. During the last century, a subset of AI called machine learning (ML) emerged. With the help of training data, machine learning algorithms build models that are capable of making intelligent decisions or predictions without being explicitly programmed [2]. Meanwhile, there are three types of learning problems in ML: supervised, unsupervised and reinforcement learning. Supervised learning can be further divided into classification and regression, and the difference between them is that classification predicts a discrete target variable and regression predicts a continuous quantity [3].

A major challenge which has affected the utilization of machine learning in some domains is the class imbalance problem [4], [5]. The class imbalance problem implies there is an uneven distribution of classes in the training data. In recent years, there has been increased interest in this problem within the machine learning community [6], and several methods have been proposed to solve the problem [7]–[9]. This classification problem exists in numerous domains, including medical diagnosis, fraud detection, and credit risk prediction.

The class imbalance problem is a challenge for predictive analytics since most ML classification algorithms were built with the premise of an equal number of instances for each class. Hence, when these algorithms are trained with imbalanced data, it leads to poor prediction results, and the performance is biased towards the majority class [8]. In literature, several sampling techniques have been often used to convert imbalanced data to balanced data, which usually lead to better performance when applied for machine learning. The main sampling methods are oversampling and undersampling [10]. The former replicates instances of the minority class to obtain a balanced dataset, while the latter reduces instances of the majority class in order to

UNIVERSITY OF JOHANNESBURG

balance the dataset [11]. However, numerous research works have proposed techniques to solve the class imbalance problem without altering the data; these techniques include ensemble learning and feature learning [12]–[16].

In this research, feature learning methods are combined with classifiers in order to improve the prediction performance of the latter on an imbalanced class problem, which is necessary because the success of traditional ML algorithms depends on the representation of the training data. Feature learning or representation learning involves the use of ML to map the original features of the input data into a new feature space, with the goal of enhancing the classification performance [16]. Deep neural networks (DNN) have been widely used for diverse feature learning tasks, and in this research, the focus is on DNN based feature learning. Figure 1.1 illustrates the ML workflow combining feature learning and a classifier. Also, there are numerous publicly available datasets for predictive analytics, and in this research, a few disease datasets and default of credit card clients dataset are considered for the experiments because they are imbalanced.

Figure 1.1: A simple ML workflow combining feature learning and classification

The methods implemented in this dissertation includes a technique for the prediction of diverse diseases, which integrates an enhanced sparse autoencoder (SAE) and softmax regression [17]. The second method involves the prediction of credit risk using a stacked sparse autoencoder (SSAE) and numerous classifiers. The methods are then contrasted with other classifiers, including classification and regression tree (CART), logistic regression, k-nearest neighbor (KNN), support vector machine (SVM), linear discriminant analysis (LDA), and conventional softmax classifier.

1.2. Problem Statement

Efficient machine learning using imbalanced data is a crucial research area, because most real-world data are imbalanced, such as medical datasets and default of credit card clients data. When using these data for machine learning, it usually leads to models which are biased towards the

majority class, and in some situations, the models completely ignore the minority class [16]. In the past, conventional ML algorithms have been used to study the class imbalance problem. Even with the advancements in deep learning, coupled with its growing popularity, not many research works have used deep learning to solve the class imbalance problem. With the state-of-the-art performance it has achieved in numerous complex problems, it could be beneficial to study the application of DNNs in solving the class imbalance problem.

Furthermore, research has shown that machine learning algorithms tend to achieve better performance when they are trained with relevant data. Hence, when building ML models, a large amount of time is spent on studying, cleaning, and preprocessing the data to ensure the most relevant features are used for training. Feature engineering and feature learning are two methods used to obtain the most relevant data for machine learning. The former is reliant on domain knowledge, and it is computationally expensive [18]. Recent research has focused on feature learning, which has resulted in improved classification results [19]–[22]. Therefore, this research aims to study and implement feature learning methods to improve the performance of ML classifiers when trained with imbalanced data. Also, the performance of the techniques implemented in this dissertation is then contrasted with other traditional ML classifiers and recent scholarly works.

1.3. Research Question and Research Objectives

The research question being considered in this work is "can effective feature learning enhance the performance of machine learning algorithms in situations where the training dataset is imbalanced?". To answer this question, the research aims to study and develop robust machine learning methods capable of obtaining excellent performance when trained with imbalanced data. The objectives of this dissertation are:

- To survey available research works on machine learning to better understand its effectiveness in the prediction of credit risk and medical diagnosis.
- To implement an unsupervised feature learning method using stacked sparse autoencoder and study its effect on the classification performance of some classifiers, with application to the prediction of credit card default.

- To develop a robust method for the prediction of diverse diseases by integrating an enhanced sparse autoencoder and softmax regression.

1.4. Contributions of the Research

The contributions of this research are outlined below:

- An extensive survey of recent research works on machine learning with application to the prediction of credit risk and medical diagnosis. The classifiers are then used to perform a comparative study with the techniques implemented in this dissertation.
- The implementation of an unsupervised feature learning method using stacked sparse autoencoder that was combined with some classifiers for the prediction of credit card default. By stacking multiple sparse autoencoders, better feature learning was obtained. Also, batch normalization was introduced to the network to address the problem of internal covariate shift which usually occurs in DNNs.
- The design of an enhanced sparse autoencoder which was combined with softmax regression for the prediction of several diseases.

1.5. Structure of the Report

UNIVERSITY OF JOHANNESBURG

The remaining part of this dissertation is structured as follows:

- Chapter 2 provides the literature review, which includes background information on machine learning and deep learning. This chapter also provides a systematic analysis of selected research works and a detailed mathematical overview of the algorithms used throughout this research.
- Chapter 3 discusses the research methodology and presents a detailed outline of the datasets utilized in training the algorithms. Also, this chapter presents an explanation of the various performance evaluation indices utilized in the course of the research.
- In Chapter 4, an implementation of a stacked sparse autoencoder is provided, which is combined with some classifiers for the prediction of credit card default.
- In Chapter 5, a method is presented, which integrates an enhanced sparse autoencoder with softmax regression for the prediction of three diseases.

- Chapter 6 concludes this dissertation and provides a summary of important findings from the research work. Chapter 6 further discusses future research direction.

1.6. Conclusion

This chapter has provided a background to the dissertation; specifically, the idea of imbalanced data and its challenge to machine learning algorithms were discussed. The chapter has presented the need to employ efficient feature learning techniques to solve the class imbalance problem. The objectives of the research and contributions of the dissertation were also presented.

UNIVERSITY
OF
JOHANNESBURG

CHAPTER 2

LITERATURE REVIEW

2.1. Introduction

Presently, artificial intelligence (AI) is transforming several sectors, including the banking industry and medical diagnosis, and has demonstrated its effectiveness in solving complex problems [23]. The introduction of machine learning (ML) has further enhanced the growth of AI [24]. ML is a subset of AI, and the term is used to imply both the academic field and the group of algorithms applied in the field. In recent times, ML has been seen as the key to the progress made in AI and has been used both in industry and academia to build models powerful enough to make accurate predictions in very complex applications [25].

The numerous achievements of machine learning can be further studied and improved for credit risk predictions and medical diagnosis, where the datasets are mostly imbalanced. In this chapter, a review of several machine learning applications of these two domains are presented. Also, an overview of the two main categories of machine learning is discussed, that is, supervised and unsupervised machine learning. Furthermore, this chapter also presents a mathematical description of the algorithms applied throughout the dissertation.

UNIVERSITY OF JOHANNESBURG

2.2. Supervised Learning

Supervised learning is considered as the most frequently used type of machine learning [26]. It can be defined as the type of machine learning where models are trained using data with known target variables. In supervised learning, the target variable can be discrete or continuous. When the target variable is a discrete value, it is termed as classification: for example, the prediction of a credit applicant as being creditworthy or not creditworthy, good or bad client, or the prediction of the absence or presence of a disease [25]. Classification algorithms include logistic regression, naïve Bayes, random forest, support vector machines, and neural networks.

When the target variable is continuous, then the supervised learning method is termed as regression. Regression methods make predictions on response variables that are continuous-valued based on what the model learned in the course of the training. Some regression algorithms include linear regression, multivariate regression, and lasso regression. The data attributes, response variables, the nature and shape of the regression curve often determine the type of regression analysis to be done. The regression curve demonstrates the correlation between the predicted and the predictor variables [3].

2.3. Unsupervised Learning

Unsupervised learning is the type of machine learning in which the algorithms obtain inference from the data without class labels. This type of machine learning is mainly utilized to obtain undefined patterns in the data [27]. Clustering is a widely used unsupervised learning method and is utilized in exploratory data analysis to extract hidden patterns from data. Other methods that fall under this category are principal component analysis and autoencoders. Furthermore, there exists another category of machine learning termed semi-supervised learning, where the algorithms are trained with some labelled data and a considerable amount of unlabeled data. Semi-supervised learning falls at the intersection between supervised and unsupervised learning [28].

UNIVERSITY OF JOHANNESBURG

2.4. Machine Learning Applications for the Prediction of Credit Card Default and Medical Diagnosis

There is a tremendous amount of research work that has applied machine learning algorithms for credit risk prediction and medical diagnosis. This section provides a general review of some of those related works. Specifically, this section presents a survey of credit risk and medical diagnosis prediction models that used the following algorithms: logistic regression, decision trees, support vector machines (SVM), k-nearest neighbors (KNN), naïve Bayes, and feature learning methods. Furthermore, Sections 4.2 and 5.2 also presents some research works particularly relevant to Chapters 4 and 5, respectively.

2.4.1. Applications of Logistic Regression

Over the years, logistic regression has been widely applied for various prediction tasks. Defaulting on credit card and loan payments is a burden on financial institutions, and machine learning algorithms such as logistic regression has been applied to predict potential defaulters, thereby enabling the lender to decline such applications. A major challenge in the prediction of credit card defaulters is that the datasets are mostly imbalanced. In [29], a method was developed to predict potential credit card defaulters using logistic regression, KNN, and naïve Bayes. Since the dataset is imbalanced, the authors performed data preprocessing using the random undersampling method, which is necessary for the classifiers to obtain good performance. To compare the performance of the various algorithms, the following metrics were utilized: accuracy, precision, true negative rate, recall, and F1 score. From the experimental results, the logistic regression model obtained better performance than the other two models with an accuracy of 95% while KNN and naïve Bayes obtained accuracies of 91% and 75%, respectively. Also, improved performance was observed after the data was undersampled before using it for training the ML algorithms.

In similar research, a comparative study was conducted using logistic regression, naïve Bayes, SVM, and KNN on an imbalanced credit card dataset [30]. Performance metrics such as accuracy, recall, precision, and true negative rate were used to carry out the comparison. From the experimental results reported, logistic regression, naïve Bayes, SVM, and KNN obtained accuracies of 99.07%, 95.98%, 97.53%, and 96.91% respectively. The results showed that the logistic regression model achieved better performance than the other models. In [31] an analysis of credit card default prediction models was carried out. The methods considered include logistic regression, multilayer perceptron (MLP), naïve Bayes, classification and regression tree (CART), and KNN. The MLP obtained the best performance, whereas logistic regression was successful in detecting important features that influence the prediction of whether a client is capable of making payment or not.

Logistic regression has also been widely used for the prediction of medical diagnosis [32]. Recently, a method was proposed for the prediction of diabetes [33]. The approach employed principal component analysis (PCA) to enhance the prediction ability of KNN and logistic regression. PCA is a mathematical algorithm that minimizes the dimensionality of its input data

UNIVERSITY OF JOHANNESBURG

while maintaining the various variations available in the data. This reduction is achieved by recognizing directions, also termed as principal components, through which the variation in the input data is maximum. The accuracy of the logistic regression classifier was increased by 1.98% after the application of PCA on the data. Furthermore, logistic regression was utilized for predicting mortality due to sepsis [34]. The proposed approach involved the analysis of sepsis indicators using feature extraction via a latent model. The simulation results recorded showed that the approach ensured the performance of the classifier was improved. Sepsis occurs when there is a massive response to bacterial infections in the blood, and it is the leading cause of death in ICU patients [35], [36]. Hence, research on the prediction of sepsis mortally is significant.

2.4.2. Applications of Decision Trees

Decision tree-based algorithms such as CART, C4.5, random forest, etc., have been applied for credit risk prediction. Recently, a comparison of machine learning methods for credit risk prediction was conducted [37]. The methods considered include decision tree-based algorithms, SVM, and logistic regression. The experimental results obtained showed that the tree-based methods, i.e. adaptive boosting (AdaBoost) and random forest, obtained the best performance. AdaBoost is usually employed to enhance the performance of the decision trees. Also, the SVM models (linear and nonlinear kernels) displayed poor performance.

UNIVERSITY OF JOHANNESBURG

In [38], a method was developed for the prediction of credit card defaulters. The approach used the synthetic minority oversampling technique (SMOTE) for addressing the imbalanced problem in the data, which enhanced the performance of random forest and six other algorithms including KNN, SVM, ANN, rotation forest, C4.5 decision tree, and NBTree. The NBTree algorithm is a hybrid of decision tree and naïve Bayes. The experimental results showed that random forest achieved the best performance with a test accuracy of 89.01%, F1 score of 89%, and area under the receiver characteristic curve (AUC) of 0.947. Meanwhile, the imbalanced nature of credit card default datasets was recently studied [39]. The authors performed both undersampling and oversampling of the data, and noticed that the latter performed better. Also, they noted that classifiers obtain better results when trained with balanced data and achieved poor results when the data is imbalanced. The study involved training several ML algorithms, and it was observed

that the gradient boosted decision tree achieved the best performance compared to the other algorithms.

Decision tree-based algorithms have also been applied for medical diagnosis; for example, in [40] a method was proposed using C4.5 decision tree algorithm to distinguish between dengue and non-dengue fever. The classifier was 84.7% accurate in performing the given task. Also, decision tree-based algorithms such as random decision forests, single decision tree, and gradient boosting have been used for detecting breast cancer [41]. The algorithms were compared using performance metrics such as accuracy, recall, and true negative rate; and the random forest obtained superior performance.

2.4.3. Applications of Support Vector Machine

There are several applications of support vector machine (SVM) for credit risk predictions. In [42], a comparative study of SVM and logistic regression was conducted using credit data. The logistic regression obtained a test accuracy of 73% and precision of 82%, whereas the linear kernel SVM achieved a test accuracy of 86% and precision of 78%. From the recorded results, SVM achieved better performance than the logistic regression. Furthermore, SVM was used for predicting credit defaulters [43]. The study utilized six credit risk datasets to demonstrate the performance of the algorithm in diverse scenarios. When compared with CART and discriminant analysis classifiers, the SVM showed superior performance.

UNIVERSITY OF JOHANNESBURG

In medical diagnosis, SVM has been employed for the prediction of numerous diseases, for example, it was used for detecting diabetes and breast cancer [44]. The method also introduced feature adaptivity to speed up the computational time and also enhance the accuracy. The proposed algorithm had better performance in comparison with the traditional SVM. The algorithm which was called an adaptive SVM achieved excellent performance on the diabetes and breast cancer predictions, with an accuracy of 100% in both cases. Furthermore, a performance comparison of some ML algorithms was conducted using heart disease dataset [45]. The algorithms include some SVM kernels together with other machine learning algorithms, including decision tree, KNN, and an ensemble classifier. The SVM kernels considered in the study include the Gaussian, linear, radial basis function, and polynomial kernels. The

experimental results showed that the linear kernel achieved better performance with AUC of 0.97 and an accuracy of 93.1%.

2.4.4. Applications of k-Nearest Neighbors

The k-nearest neighbor (KNN) algorithms have applications in credit risk analysis. In [46], a weighted KNN (WKNN) technique was developed to predict credit risk using data from an Indonesian bank. The research evaluated some kernels and reported that the rectangular and Gaussian kernels obtained the best performance with both kernels having accuracy of over 82%. In [47], a study was conducted to analyze the default of credit card clients data; and seven machine learning algorithms were used for the analysis, including KNN, naïve Bayes, random forest, logistic regression, decision tree, and two SVM kernels. From the research, it was concluded that out of the independent variables in the data, there are just a few that can be effectively used to determine if clients would default or not.

KNN has been widely used for disease predictions [48], [49]. An approach was developed for the prediction of heart disease by combining genetic algorithm and KNN [50]. The approach involved the ranking of attributes according to their importance and the elimination of irrelevant features using genetic search as the measure of goodness. By training the KNN with the most important features, an improved performance was observed. In [51], another method was developed to predict heart disease using the ant colony optimization technique to perform feature selection, and a hybrid KNN classifier performed the prediction. The proposed approach obtained a classification accuracy of 99.2%, which showed superior performance when compared with some machine learning classifiers such as decision tree, naïve Bayes, SVM, and traditional KNN. In another research, KNN was utilized to predict whether patients with prediabetes have a two-year risk of developing type 2 diabetes mellitus [52]. The dataset used for training the algorithm contained 1647 samples, with features from clinical and laboratory tests. The KNN classifier achieved a test accuracy of 96%, a true negative rate of 78%, and true positive rate of 99%.

UNIVERSITY OF JOHANNESBURG

2.4.5. Applications of Naïve Bayes

Naïve Bayes (NB) classifiers have been used to understand credit data and make appropriate predictions. Recently, a study was conducted to predict the probability of clients defaulting their credits [53]. The study utilized credit data from a bank in Tunisia containing 924 samples, and the classification was performed using the NB machine learning algorithm. From the experimental results, the algorithm obtained a classification accuracy of 63.85%. Furthermore, the study reported that the probability of a client defaulting on the credit was better represented by some variables, including solvency, leverage, profitability, working capital, and cash flow measures.

Also, naïve Bayes classifiers have been used for prediction of medical diagnosis. In [54], a Gaussian NB was utilized for predicting lung and breast cancers with test accuracy of 90% and 98%, respectively. Furthermore, naïve Bayes was used to classify melanoma (skin cancer) as either malignant or benign [55]. The study used images from an epiluminescence microscopy for training the model. The NB classifier showed superior performance when compared with a decision tree, with the former having an accuracy of 98.8% and the latter obtained an accuracy of 92.86%. In another research, a hidden naïve Bayes (HNB) was proposed to identify heart disease [56]. The difference between the HNB and traditional naïve Bayes is that the HNM modifies the independence assumptions between the predictor variables that exist in the traditional naïve Bayes algorithm. The HNB classifier achieved a test accuracy of 100%.

UNIVERSITY OF JOHANNESBURG

2.4.6. Applications of Feature Learning

Feature learning techniques have been used in both the prediction of credit card default and medical diagnosis. Feature learning is used mainly to map high dimensional input data to low dimension for effective classification [57]. In [58], a method was developed for the classification of skin lesions using a generative model. This was achieved by an autoencoder in which both the encoder and decoder undergo some adversarial training using different discriminator networks. The efficacy of the method was demonstrated through the classification of images from epiluminescence microscopy, and the method obtained excellent results. The application of feature learning can also be seen in the prediction of breast cancer. In [59], an autoencoder was used for detecting breast cancer. The autoencoder performed feature learning of

the common structural patterns in regular breast cancer images. After the training, the autoencoder was able to identify images that are different from the standard images.

In [60], a deep autoencoder was applied to identify different cancers. The autoencoder aimed at discovering hidden correlations in input data, thereby leading to a classification accuracy of 100% in the three datasets considered in the study. This showed the ability of deep autoencoders and the importance of feature learning. Furthermore, there are numerous feature learning research works available in the literature, including stacked sparse autoencoder to detect Parkinson's disease [61], a stacked sparse autoencoder combined with support vector machine to detect osteoporosis [62], an autoencoder based recurrent neural network for the prediction of diseases [63], and a stacked sparse autoencoder method for detecting lung cancer [64].

2.5. Overview of Machine Learning Algorithms

This section presents a detailed description of the machine learning algorithms used throughout this dissertation, and this is necessary to lay a proper foundation for the research since these algorithms will be benchmarked against the methods developed in this dissertation. In the following subsections, brief and concise explanations of the algorithms are provided, including their mathematical overview.

UNIVERSITY OF JOHANNESBURG

2.5.1. Logistic Regression

Logistic regression is a statistical model employed for analysing data that contains more than one predictor variable to obtain the predicted variable. In logistic regression, the predicted or response variable is usually binary [8]. Logistic regression is well suited for credit risk prediction and medical diagnosis since their class attributes are generally binary. Also, this algorithm aims to obtain a model with the best fit line that describes the relationship between the target variable and the predictor variables, and it generates the variables expressed as:

$$logit\ (p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k \tag{2.1}$$

where p denotes the probability of presence of the attribute of interest. This detects a logit transformation of the likelihood of the presence of the attribute of interest. Secondly, the logit transformation is illustrated as the logged odds shown as:

$$odds = \frac{p}{1-p} = \frac{probabilty\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characterristic} \tag{2.2}$$

$$logit(p) = ln\left(\frac{p}{1-p}\right) \tag{2.3}$$

Another variant of logistic regression is the softmax regression, also called multinomial logistic regression, and it is used to develop models where the data has multiple class variables. The softmax function is represented as:

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}} \ (i = 1,2, \dots, N) \tag{2.4}$$

where $x_1, x_2, \dots, x_N$ represents the input values and $f(x_i)$ is the output, which is the probability that the sample belongs to the *i*-th class [65]. In the course of this research, both the softmax regression and logistic regression will be utilized.

2.5.2. Decision Tree

Decision tree algorithms are supervised machine learning methods used for classification and regression [66]. A decision tree model where the predicted variable is a set of binary values is called a classification tree. In contrast, when the predicted variable is a continuous value, it is called a regression tree. Tree-based algorithms are often used in several applications because of their simplicity [67]. Decision trees contain a root node, leaf nodes, and branches, which are the three main parts. The tree-building process starts from the root node, both the root node and leaves comprise of questions or criteria that should be met. The branches are arrows that connect the nodes, and they show the flow from questions to answer. There are many tree-based machine learning algorithms such as Classification and regression tree (CART) [68], Iterative Dichotomiser (ID3) [69], and C4.5 [70]. In this dissertation, the CART algorithm is utilized, and it uses the Gini index to obtain the probability that a given variable is incorrectly classified when it is randomly selected [71]. To compute the Gini index for a sample data having J classes assuming $i \in \{1,2, \dots, J\}$:

$$Gini = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{J} p_i \, (1 - p_i) = \sum_{i=1}^{J} (p_i - p_i^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2 = 1 - \sum_{i=1}^{J} p_i^2 \tag{2.5}$$

where p_i denotes the probability that an instance is classified into a specific class [72].

UNIVERSITY OF JOHANNESBURG

2.5.3. Support Vector Machine

Support vector machine (SVM) is a machine learning algorithm which can be used to solve regression and classification tasks. It is based on statistical learning theory and has been used to make accurate predictions in different fields [73]. SVM has been applied to linear classification tasks, and it is also useful to solve non-linear classification problems using a technique called kernel trick [74], which transforms a non-linear separable input into a high dimensional space at which point a hyperplane capable of separating the data is established. There are multiple SVM kernels such as polynomial, radial basis, linear, Gaussian, and nonlinear kernels.

Supposing the input data is $T = \{(x_i, y_i)_N\}$, in which $y_i \in \{+1, -1\}$, the goal of the SVM classifier is to get a hyperplane capable of dividing the space into a pair of spaces corresponding to the classes in the input data [75]. A hyperplane here implies a linear function of $x, f(x) = \langle w.x \rangle + b$, in which

$$y_i\big(f(x)\big) = y_i(\langle w.x \rangle + b) > 0 \tag{2.6}$$

where w denotes a weight vector, b represents bias, whose value is a scalar quantity. Therefore, we represent the separating hyperplane as:

$$f(x) = \langle w.x \rangle + b = 0 \tag{2.7}$$

UNIVERSITY OF JOHANNESBURG

The generalization ability of SVM is outstanding because the generalization error is minimized while the separating margin is maximized by the algorithm [75], which is represented and solved through constrained optimization, that is minimize $\frac{1}{2}\|w\|^2$ or maximize the margin $\frac{2}{\|w\|}$ with respect to $y_i(\langle w.x \rangle + b) \geq 1$. The Lagrange multipliers strategy is used to solve this constrained optimization task. After computing the Lagrange L and including an unknown scalar α, the following is achieved:

$$w = \sum_{i=1}^{N} \propto_i y_i x_i \tag{2.8}$$

$$L = \sum_{i=1}^{N} \propto_i - \frac{1}{2} \sum_{i,j=1}^{N} \propto_i \propto_j y_i y_j x_i . x_j \tag{2.9}$$

The Lagrangian L is maximized to obtain the coefficients of $\propto_i$, with respect to the constraint:

$$\sum_{i=1}^{N} \propto_i y_i = 0 , \propto_i > 0 \quad (2.10)$$

After obtaining the coefficients of $\propto_i$, a hypothesis is obtained, corresponding to a linear combination of the input data points. Lastly, the decision function is expressed according to:

$$h(x) = sgn(\langle w.x \rangle + b) = sgn\left(\langle \sum_{j=1}^{N} \propto_j y_j x_j . x \rangle + b\right) \quad (2.11)$$

From Equation (2.11), it is observed that SVM learning depends on the dot products of input pairs, whereas prediction of unseen sample entirely depends on the dot product of the sample under consideration with the input or training data [75]. Furthermore, SVM is suitable for applications where the dataset is small, and when the dataset is increased, the performance of the algorithm becomes poor.

2.5.4. k-Nearest Neighbors

K-nearest neighbors (KNN) is a machine learning algorithm which has the ability to perform classification and regression. However, it is mostly used for classification, and it is considered as a lazy learning and non-parametric algorithm [76]. It is non-parametric for the reason that it does not make any assumption about its input data, and lazy learning means KNN generalizes the data following a query [77]. Furthermore, KNN performs classification on unlabeled samples by placing them in the class of correlated labelled samples with regard to similarity. Meanwhile, there exists numerous means to conduct the KNN computations such as Hamming, Manhattan, and Euclidean distance. The Euclidean distance is often utilized for most applications [78], and can be represented mathematically as:

$$D(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \quad (2.12)$$

From Equation (2.12), p and q are samples that are being compared having n features. When applying the KNN algorithm, the value of k has to be selected, and it represents the nearest data points or the number of neighbors [78]. This algorithm is quite simple to implement and has been used in several areas, including prediction of credit risk and medical diagnosis.

UNIVERSITY OF JOHANNESBURG

2.5.5. Naïve Bayes Classifier

Naïve Bayes classifier is developed based on Bayes' theorem. The algorithm is called naïve because it assumes that the input features are independent of each other [79]. There exists diverse naïve Bayes classifiers such as multinomial and Gaussian naïve Bayes and they are mainly used in situations where the dataset is large [80]. According to Bayes' theorem, a class variable (c) of a sample data (x) is obtained by computing the posterior probability of the value $\mathrm{P}(c|x)$ as:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{2.13}$$

where $P(x|c)$ represents the posterior probability of sample data x conditioned on class c, and $P(c)$ denotes the prior probability of class variable c. $P(x)$ represents the prior probability of sample data x.

2.5.6. Feature Learning

Feature learning, also called representation learning, comprises of techniques which allow machines to automatically extract the representations required to detect features or to classify raw data [26]. Unsupervised feature learning discovers or learns the features from data that is not labelled. With this method, there is minimal need for non-automatic feature engineering. Supervised feature learning methods include multilayer perceptron, supervised dictionary learning, and neural networks. Meanwhile, unsupervised feature learning methods include autoencoders, matrix factorization, independent component analysis, and several clustering techniques. Effective representation learning can simplify classification problems, and this is a vital step in many domains, especially in the prediction of medical diagnosis and credit risk predictions [81]. In the next subsection, a brief discussion of the autoencoder is presented since it is the feature learning method under consideration in this dissertation.

2.5.7. Autoencoders

Autoencoder is a neural network capable of learning representations or features automatically when given some training data, thereby making it a perfect method for removing the complexities associated with manual feature engineering when performing ML tasks. Autoencoders are suitable for denoising data, dimensionality reduction, and pre-training DNNs

UNIVERSITY OF JOHANNESBURG

[82], and they consists of two parts: namely, encoder and decoder. The main function of autoencoders is to reconstruct an input data at the output. The encoder is used to extract features from the training data to obtain the hidden layer via some nonlinear mappings. At the same time, the decoder predicts the output vector to reconstruct the original input vector. By imposing some constraints on the network, the autoencoder can discover excellent representation of the training data [83]. There exists several categories of autoencoders, such as sparse, convolutional, variational, and denoising autoencoders. However, this dissertation focuses on the sparse autoencoder, and its mathematical representation is presented in Chapters 4 and 5.

2.6. Research Gap

When building machine learning models, an enormous amount of labelled data is usually required. The use of feature engineering to obtain features from raw data could be expensive because domain knowledge is needed, and it is also time-consuming. Unlabeled data is easily accessible in the financial and health sectors. Additionally, current research works have shown the effectiveness of feature learning techniques in developing models using unlabeled data, thus minimizing the use of labelled data [84]–[88]. Several research works have applied machine learning algorithms for prediction of credit risk and medical diagnosis, but not many employed feature learning to select the most relevant features to train the classifiers, which can improve the classifier performance, especially when dealing with the class imbalance problem. Motivated by several advances in feature learning, this research aims to build on what has been done and develop improved machine learning methods based on unsupervised feature learning for the prediction of credit risk and medical diagnosis, thereby minimizing the over-reliance on feature engineering in these crucial domains.

Furthermore, one vital problem when applying machine learning in these domains is that the data is usually imbalanced, that is, the negative samples exceed the positive samples, and this leads to poor performance by the ML classifiers [89]. Feature learning has been studied to solve the imbalanced class problem, and this is achieved because the learned representations usually amplify attributes of the input that is vital for discrimination, while also suppressing attributes that are irrelevant [16]. The feature learning methods developed in this dissertation ensures the classifiers are trained with the most relevant data. Therefore, the classifiers are more robust when handling datasets that are not balanced.

UNIVERSITY OF JOHANNESBURG

2.7. Conclusion

In this chapter, a general summary of machine learning and its algorithms is presented, and a review of some recent research works that applied ML for the prediction of credit risk and medical diagnosis is also is provided. The chapter also discussed the mathematical representation of the ML algorithms that are used in the course of this research. The content of this chapter is vital as it provides the required details of the numerous methods employed during the study. Furthermore, the research gap that the dissertation aims to fill is also discussed. Lastly, the methods proposed in this dissertation will be benchmarked against the machine learning methods presented in this chapter.

UNIVERSITY OF JOHANNESBURG

CHAPTER 3

RESEARCH METHODOLOGY

3.1. Introduction

In this chapter, the methodology used for the dissertation is presented, and a detailed description of the various datasets used in training the models. The chapter also provides information regarding the various performance evaluation metrics used to assess the performance of the models. Meanwhile, the experimental research approach is utilized for this research. This approach is a scientific research method in which some predictor variables are manipulated which have an impact on the predicted variables. The effect of the predictor variables on the predicted variables is essentially measured to aid scientists in deriving the necessary inference from the data [90].

3.2. Credit Card Dataset

The financial sector is one of the areas with highly imbalanced data such as credit card datasets. Therefore, in this work, the default of credit card clients dataset [91] is utilized for training and testing one of the proposed methods. This dataset comprises of 30,000 instances and 25 attributes, which includes demographic and financial records. The dataset was obtained from the University of California, Irvine (UCI) ML repository, and it was established to predict customers that are likely to default on their credit card payments in Taiwan. From the 30,000 instances, 23,364 are non-default, and 6,636 are default cases, which shows that the datasets is quite imbalanced.

3.3. Heart Disease Datasets

Heart diseases are considered to be among the most dangerous diseases affecting humans in their middle and old ages, and it affects men more than women. Also, research has shown that this disease constitute one-third of all deaths globally. Over 17 million people die of heart-related diseases each year. Some of the risk factors associated with the disease include poor diet, obesity, diabetes, family history and age [92]. Numerous machine learning models have been

developed to predict heart diseases [93]–[96]. There are several heart disease datasets available in online repositories. In this research, the Framingham heart disease dataset is used. This dataset was obtained from the Kaggle website [97], and it was established after a cardiovascular study on residents of Framingham, Massachusetts, which aimed to predict patients' 10-year risk of developing heart disease. The dataset comprise of 4238 instances and 16 attributes. The attributes include demographic, behavioral, and medical risk factors. From the 4238 instances, 3594 are negative, while 644 are positive. This dataset is also imbalanced.

3.4. Cervical Cancer Dataset

A common disease that affects women globally is cervical cancer: a disease caused by the human papillomavirus (HPV). This type of cancer forms in the tissues of the cervix. The risk of developing cervical cancer could be minimized through screening for the disease, and taking the vaccine, which protects against HPV [98]. Early detection and treatment reduces the spread of cervical cancer and increase the chances of survival [99]. In this research, we employ the cervical cancer (risk factors) dataset [100] for training and testing one of the proposed methods. The dataset was obtained from the UCI machine learning repository, and it has been widely utilized for studying and developing machine learning models to predict cervical cancer [101]–[103]. The dataset comprises of 858 samples and 32 attributes together with four classes: Hinselmann, Schiller, Cytology, and Biopsy, which represents the four tests usually carried out to detect cervical cancer. Meanwhile, the most accurate test for detecting this disease is the biopsy [102], and it is mostly used as the predicted variable. Among the 858 samples, 803 are negative, and 55 are positive, which shows the dataset is highly imbalanced.

UNIVERSITY OF JOHANNESBURG

3.5. Chronic Kidney Disease Dataset

Chronic Kidney Disease (CKD) is the gradual loss of kidney function. More than 10% of the global population and 15% of South Africa's population suffer from this disease. The major causes of chronic kidney disease include diabetes, high blood pressure, obesity, heart disease, and family history [104]. The CKD dataset [105] used in this research was obtained from the UCI machine learning repository, and it consists of 400 samples with 250 positive cases and 150 negative cases. The dataset is fairly balanced compared to the credit card, heart disease, and cervical cancer datasets.

3.6. Performance Evaluation Metrics

There are several metrics used to test the performance of machine learning algorithms. This section discusses the metrics utilized to evaluate the performance of the proposed methods. Accuracy is the most widely used performance evaluation metric when dealing with classification problems [106]. It is a statistical measure and is defined as the ratio of correct predictions (i.e. true negative (TN) and true positive (TP)) made by the classifier divided by the total predictions made, including false negatives (FN) and false positives (FP). Precision is the ratio of correct positive predictions to the total positive instances. High precision demonstrates low false-positive rate. The next metric is sensitivity, also known as recall; it is the ratio of correctly predicted positive instances to all the actual positive instances. The F1 score can be defined as the weighted average of precision and recall. Hence, this metric takes into account both false negatives and false positives. The F1 score is significant when analyzing the performance of algorithms trained with imbalanced datasets. These performance metrics can be represented mathematically as:

$$Classification\ accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.1}$$

$$Precision = \frac{TP}{TP+FP} \tag{3.2}$$

$$Recall\ = \frac{TP}{TP+FN} \tag{3.3}$$

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} \tag{3.4}$$

UNIVERSITY OF JOHANNESBURG

Table 3.1 shows a confusion matrix, which visualizes the relationship between the variables *TP*, *FP*, *TN*, and *FN*.

Table 3.1: Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	*TP*	*FN*
	Negative	*FP*	*TN*

Furthermore, the receiver operating characteristic (ROC) curve and the AUC are also used in this research. The ROC curve is a graphical plot that shows the diagnostic ability of a classification

model, and it is plotted with true positive rate (TPR) against the false positive rate (FPR). AUC demonstrates the ability of a classifier to distinguish the various classes. High AUC means the model predicts negative classes as negative and positive classes as positive, and the AUC values usually start from 0 to 1, where 0 implies a totally inaccurate model, and a value of 1 means a fully accurate model [107], [108].

3.7. Experimental Environment

In order to build the various ML models, the Python programming language is used. Specific machine learning libraries used include Scikit Learn (also called sklearn) and Keras. Furthermore, several libraries are also used during the model development, data manipulation, and evaluation of the models, including Numpy, Pandas, and Matplotlib. The simulations were performed using an Intel Core i5-6300U computer running at 2.40 GHz, with 16GB RAM.

3.8. Conclusion

In this chapter, the research methodology adopted for the dissertation has been discussed. The datasets used in developing the models were also presented and discussed. And lastly, the chapter gave a brief description of the various performance indices used to evaluate the effectiveness of the proposed methods. In the following chapters, the enhanced machine learning methods for prediction of credit risk and medical diagnosis are presented.

UNIVERSITY OF JOHANNESBURG

CHAPTER 4

AN IMPROVED MACHINE LEARNING APPROACH FOR PREDICTION OF CREDIT CARD DEFAULT

4.1. Introduction

In artificial intelligence (AI) and machine learning (ML) tasks such as classification and clustering, the input data tends to influence the performance of the algorithms. Optimal performance is obtained when algorithms are given suitable data. To this end, some ML methods focus on processing high dimensional data, including linear dimensionality reduction methods such as linear discriminant analysis, principal component analysis, and multiple dimensional scaling and nonlinear dimensionality reduction techniques such as isometric mapping and Laplacian Eigenmap. Meanwhile, feature engineering and representation learning are the two main methods used to achieve representation from raw data. Recent research has focused on the latter since feature engineering methods are usually dependent on domain knowledge, are labor-intensive, and time-consuming [109]. Furthermore, representation learning methods tend to learn a representation from data automatically, which can then be used for classification. An autoencoder (AE) is a type of unsupervised representation learning.

Autoencoders are unsupervised neural networks having multiple layers, including input, hidden, and output layers [110]. Autoencoders tend to learn a representation of the input data, usually for dimensionality reduction, through training the network to disregard noise. Also, the AE attempts to create a representation of the initial input [111], [112]. There are different types of autoencoders, including sparse, denoising, contractive, variational, and convolutional autoencoders [113].

Credit card default/fraud detection is a crucial problem that has gotten the attention of machine learning researchers, and a significant number of approaches have been proposed [114]–[117]. However, the problem is still challenging since most credit card data seem to suffer from class imbalance, as non-fraud transactions overwhelmingly supersede fraud transactions, making it difficult for many machine learning algorithms to achieve good performance. Meanwhile, a good feature representation can be obtained from the dataset, which can enhance the classification

performance of the algorithms. Representation learning is a possible solution to the challenge of credit card default and fraud prediction because of its remarkable feature learning ability in large and imbalanced data.

Basic autoencoders aim at learning a representation of the data and reconstructing the output as close as possible to the input data, however, training the autoencoder network in such a way that encourages sparsity can result in better feature learning. Sparsity induced neural networks have been extensively applied in image recognition and several other applications resulting in state-of-the-art performances [118]–[120].

In this chapter, an approach is presented to improve the classification performance of classifiers by using the unsupervised feature learning capability of autoencoders. During the training of the autoencoder, sparsity is encouraged, and the model is optimized using the AdaMax algorithm [121] instead of the conventional stochastic gradient descent. To ensure accurate feature representation, multiple sparse autoencoders are stacked to get the final model. Also, to further prevent overfitting and enhance the performance, speed, and stability of the network, we introduced the batch normalization technique [122] to the network. The low-dimensional features were then used to train the various classifiers, including logistic regression (LR), support vector machine (SVM), classification and regression tree (CART), *k*-nearest neighbor (KNN), and linear discriminant analysis (LDA). The performance of these classifiers is then benchmarked against an instance where the classifiers were trained with the raw data. Further comparison is made with other scholarly works, and our proposed method shows better performance. The main contributions of this study can be summarized as follows:

- To construct an effective artificial neural network for feature learning using stacked sparse autoencoder.
- To improve the classification performance of various classifiers using the proposed stacked sparse autoencoder.
- To demonstrate the effectiveness of feature learning on the performance of classifiers using the credit card dataset.

The rest of the chapter is structured as follows. Section 4.2 provides a review of related works that utilized different types of autoencoders. The proposed methodology is presented in Section 4.3, while Section 4.4 presents a case study of credit card defaulting prediction models. The

obtained results are presented and discussed in Section 4.5. Lastly, Section 4.6 concludes the research and highlights a few future research directions.

4.2. Related works

Several applications of autoencoders exist in literature, and they achieved excellent performance. This section discusses some previous works that utilized various autoencoders and lay the foundation for the proposed stacked sparse autoencoder network. Sun et al. [123] proposed a method for fault diagnosis by applying a sparse stacked denoising autoencoder due to its robustness and data reconstruction capability, which improved the diagnostic accuracy. The autoencoder was used together with an optimized transfer learning algorithm. Similarly, Zhu et al. [124] proposed a novel stacked pruning sparse denoising autoencoder. To effectively train the autoencoder, a pruning function was introduced to the architecture in order to restrict suboptimal features from being part of the final model. The method was utilized for detecting faults in rolling bearings, and when compared with other fault diagnostic models, their approach showed superior performance.

Furthermore, Sankaran et al. [125] proposed a feature extraction method using an autoencoder network, and $\ell_{2,1}$-norm based regularization was used to achieve sparsity. The authors identified that due to the many training variables that exist, several representation learning methods are susceptible to overfitting, which was mitigated in their approach due to the regularization technique. The performance of their model was studied on publicly available latent fingerprint datasets, and it achieved an improved performance. Chen et al. [22] developed a technique to solve the problem of computational complexity in deep neural networks. The technique used a sparse autoencoder (SAE) for learning facial features and softmax regression to classify expression features; the softmax regression aimed at handling multiple data at the SAE output. Also, the problem of local extrema and the challenge of gradient diffusion during training was handled when the network weights were fine-tuned, and this improved the performance of the architecture.

Most approaches used to implement autoencoders depend on the single autoencoder model, and this presents a problem when learning different characteristics of data. Yang et al. [21] developed an approach to solve the problem by implementing a feature learning framework

UNIVERSITY OF JOHANNESBURG

using serial autoencoders. The technique achieved superior representation learning by connecting two distinct autoencoders serially. When compared to baseline methods, the proposed approach showed significant improvement. Meanwhile, Al-Hmouz et al. [126], introduced a logic-driven autoencoder, whereby the network structure was achieved using some fuzzy logic operations. The autoencoder was also optimized using gradient-based learning.

Furthermore, Musafer et al. [127] proposed a type of sparse autoencoder in which sparse regularization was imposed on the weights. The method was combined with random forest and applied to a network intrusion detection system. Lastly, sparse autoencoder networks have achieved remarkable performance in representation learning [128], [129]. However, better representation learning can be gotten when multiple sparse autoencoders are stacked and optimized effectively, which is the focus of this research.

4.3. Proposed Methodology

In this section, we present the step by step method applied to develop the proposed autoencoder. Autoencoders generally contain two functions, i.e., the encoder and decoder [130]. Assuming the original input is x, the autoencoder encodes it into a hidden layer h so as to decrease the dimension of the input, which is subsequently decoded at the output. The encoding process of the input vector can be mathematically represented as:

$$h = \sigma(Wx + b) \tag{4.1}$$

where σ represents the activation function, W and b denote the weight and bias matrices, respectively. Subsequently, the hidden representation is decoded to obtain data that is as near as possible to the original data x, and this process is represented as:

$$\hat{x} = \sigma(W'h + b') \tag{4.2}$$

The sigmoid function [96], which is used as the activation function in this work is described as

$$\sigma = \frac{1}{1+e^{-x}} \tag{4.3}$$

The disparity between the original input x and the reconstructed input $\hat{x}$ is called reconstruction error. We utilize the mean squared error (MSE) function to optimize the weights and bias parameters. The MSE function is represented as:

UNIVERSITY OF JOHANNESBURG

$$E = \frac{1}{N}\sum_{i=1}^{N}\|\hat{x}_i - x_i\|^2 \tag{4.4}$$

In the network hidden layer, the average activation of nodes is represented as:

$$\widehat{p_j} = \frac{1}{N}\sum_{i=1}^{N} h_j(x_i) \tag{4.5}$$

To induce sparsity in the autoencoder, we limit $\widehat{\rho_j} = \rho$, where ρ is the sparsity proportion, and it is usually a small positive value close to 0. Therefore, the Kullback–Leibler (KL) divergence between $\widehat{\rho_j}$ and ρ is minimized according to:

$$KL(\rho||\hat{\rho}) = \rho\log\left(\frac{\rho}{\widehat{\rho_j}}\right) + (1-\rho)\log\left(\frac{1-\rho}{1-\widehat{\rho_j}}\right) \tag{4.6}$$

Also, to ensure better feature representation and, by extension, enhance the performance of the classifiers, multiple sparse autoencoders are stacked. A stacked sparse autoencoder (SSAE) can comprise of numerous sparse autoencoders whereby the output of every layer is connected to the input of the next layer [131]. The SSAE is based on research conducted by Hinton and Salakhutdinov [132], where they proposed a deep neural network with layer by layer initialization. The SSAE error function is expressed as:

$$J(W,b) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}\left\|\sigma\left(Wx^{(i)}+b\right) - y^{(i)}\right\|^2\right] + \frac{\lambda}{2}\sum_{l=1}^{n-1}\sum_{i=1}^{s_l^c}\sum_{j=1}^{s_l^r}\left[W_{ji}^{(l)}\right]^2 \tag{4.7}$$

where N and n denotes the number of samples and the number of layers, respectively, the original input is x, and y denotes the corresponding label. The regularization coefficient is represented by λ. s_l^r and s_l^c denotes the rows and columns of the matrix $W_{ji}^{(l)}$ [133]. By adding the sparsity term to (4.7), the overall cost function of the SSAE becomes:

$$L_{Sparse}(W,b) = J(W,b) + \beta\sum_{j=1}^{S} KL(\rho||\hat{\rho}) \tag{4.8}$$

where S represents the entire amount of neurons in a layer and $\boldsymbol{\beta}$ represents the sparsity regularization constant; which controls the sparsity penalty term. We now have three optimization parameters, including $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\rho}$, and we set their values as 3, 0.0001, and 0.05, respectively. In the stacked sparse autoencoder network, a neuron is said to be active when its output is a value near 1, while it is inactive when the output value is closer to 0 [8]. Algorithm 4.1 shows the proposed SSAE procedure, and Figure 4.1 shows its structure. For simplicity, the decoder parts of

UNIVERSITY OF JOHANNESBURG

the various SAE are not shown. The output of the SSAE is then used to train the various classifiers.

Algorithm 4.1. Proposed methodology of the SSAE

Input:

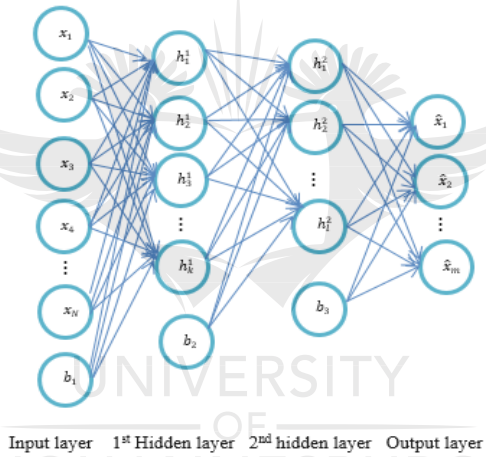
train set x

Process:

1. Start
2. Initialize σ, W, W', b, b'
3. Obtain the cost function according to (4)
4. Apply weight penalty to the cost function according to (7)
5. Add the sparsity regularizer to the cost function according to (8)
6. Train network until convergence
7. End

Output:

Reconstructed representation of the input



UNIVERSITY OF JOHANNESBURG

Figure 4.1. Structure of the proposed SSAE model

The greedy layer-wise training strategy proposed by Bengio et al. [134] is employed to successively train every layer of the SSAE in order to obtain access to the weights and bias parameters of the network. Also, the network is fine-tuned using the backpropagation algorithm to obtain the best parameter settings. The AdaMax algorithm [121], a variant of the adaptive moment estimation (Adam) algorithm that uses the infinity norm, was applied to optimize the autoencoder network. Lastly, we introduced the batch normalization technique [122] to prevent overfitting and enhance the performance, speed, and stability of the network.

4.4. Case Study of Credit Card Defaulting Prediction Models

Credit risk prediction is a crucial task in the financial sector. Most financial institutions grant loans, mortgages, and credit cards, among many other services. Due to the rising number of credit card clients, these institutions have faced an increasing default rate. They are thereby resorting to the use of machine learning methods to automate the application process and predict the probability of a client's future default. However, several machine learning methods have been developed in various literature with varying performances. A major limitation to achieving optimal performance in the credit card default prediction is that the datasets are highly imbalanced, i.e., the instances where clients do not default are more than the defaulting cases.

Certain studies have used the default of credit card clients dataset [91] and achieved good performance. For example, Prusti and Rath [135] used various algorithms such as decision tree, KNN, SVM, and multilayer perceptron to make predictions on the dataset. Additionally, they proposed a method that hybridized decision tree, SVM, and KNN, which gave improved performance compared to the stand-alone algorithms. Sayjadah et al. [136] conducted a comparative study of logistic regression, random forest, and decision tree for credit risk prediction. The experimental results showed that random forest achieved superior performance with an accuracy of about 82%.

Furthermore, because the dataset is imbalanced, a method was proposed to tackle the problem via the synthetic minority over-sampling technique (SMOTE) [38]. Applying the SMOTE method together with seven other algorithms, the random forest algorithm achieved the best performance with an accuracy of 89.01% and F1-score of 89%. Lastly, Hsu et al. [137] and Chishti and Awan [138] also proposed models to predict the defaulting of credit card clients and achieved comparable performance. However, we are aiming to improve on what has been done by applying our proposed method on the same dataset.

4.5. Results and Discussions

In this work, the defaulting of credit card client dataset [91] is used. The dataset was obtained from the University of California Irvine ML repository, and it contains 30,000 instances and 25 attributes, including demographic and financial records. The dataset was established to predict customers who are likely to default on payments in Taiwan. Out of the 30,000 instances 23,364 are non-default and 6,636 are default cases. The rationale behind the dataset is for financial

institutions to be able to identify possible customers who will default on their credit card payments, thereby declining such applications. We use the 70-30% train-test split strategy in the experiments. The SSAE is trained with the training set in an unsupervised fashion, whereas the test set is input with the learned SSAE model to get the low-dimensional data. The low-dimensional training set is then used to train the various classifiers, and the performance tested on the test set. The number of neurons in the first and second hidden layers was set at 100 and 85, respectively.

To efficiently evaluate the performance of our approach, we utilize certain performance metrics, as explained in Section 3.6. The Python programming language was utilized for the computations. In order to demonstrate the effectiveness of the proposed SSAE, five classifiers are employed. Therefore, we first show the performance of these classifiers on the raw dataset; the classifiers include CART, LR, KNN, SVM, and LDA, and the respective performances are shown in Table 4.1.

Table 4.1: Performance of the base classifiers on the dataset

Algorithm	Accuracy (%)	Precision (%)	Sensitivity (%)	F1 score (%)
LR	78	62	78	69
CART	73	74	73	73
KNN	75	71	75	72
SVM	36	74	36	35
LDA	81	79	81	78

UNIVERSITY OF JOHANNESBURG

Table 4.2: Impact of the proposed SSAE on the base classifiers

Algorithm	Accuracy (%)	Precision (%)	Sensitivity (%)	F1 score (%)
LR	90	87	91	89
CART	86	84	84	84
KNN	89	87	90	88
SVM	88	86	88	87
LDA	90	91	90	90

Table 4.3: Comparison with prior works

Literature	Accuracy (%)	Precision (%)	Sensitivity (%)	F1 score (%)
Prusti and Rath [135]	82.58	96.83	83.57	89.71
Sayjadah et al. [136]	81.81	-	-	-
Subasi and Cankurt [38]	89.01	-	-	89
Hsu et al. [137]	80.2	-	-	-
Chishti and Awan [138]	82	84	96	89
Proposed SSAE+LDA	90	91	90	90

Table 4.2 summarizes the results obtained when the classifiers were trained using the features learned by the SSAE. It is observed that the learned features improved the performance of the classifiers. Furthermore, the results show the ability of the proposed SSAE to achieve good representation learning. Also, the best performing model from our experiments, which is the LDA, is used to compare with some recent research works discussed in Section 4.4, and this is shown in Table 4.3. To give a fair comparison, we focused on studies that used similar datasets. From Table 4.3, it can be seen that the proposed technique outperformed those in the stated literature.

From the above results, the proposed approach achieved better performance compared to the other methods. The improved performance can be attributed to the proposed SSAE that was able to achieve good feature learning. Also, the results have shown the capability of deep learning in achieving exceptional performance in different tasks, including feature representation. Lastly, this study has demonstrated the importance of training machine learning algorithms with suitable data, and that improved performance can be obtained not only by hyper-parameter tuning but also and more efficiently by effective feature learning.

UNIVERSITY OF JOHANNESBURG

4.6. Conclusion

In the financial industry, accurate prediction of potential credit card defaulters is a crucial task, and several machine learning algorithms have been utilized with varying performances. In this research work, a stacked sparse autoencoder was developed to achieve excellent feature learning. In the proposed SSAE, batch normalization was introduced to enhance the performance, speed, and stability of the model, and to further prevent overfitting. Also, the model was optimized using the AdaMax algorithm. The learned data was then used to train five machine learning algorithms. When compared with a case where the algorithms were trained with the raw data, the

proposed method showed superior performance. Furthermore, the results were compared with methods in some recent literature that used a similar dataset, and the proposed approach also showed improvement. Future research will focus on studying the effect of different optimizers and stacking diverse autoencoders, and observing the resultant impact on the feature learning process. Also, future research can consider comparing the feature leaning capability of the stacked sparse autoencoder with other feature learning and feature engineering methods. Lastly, future works can further test the performance of the proposed method on other credit card datasets.

UNIVERSITY
OF
JOHANNESBURG

CHAPTER 5

A SPARSE AUTOENCODER AND SOFTMAX REGRESSION METHOD WITH APPLICATION TO MEDICAL DIAGNOSIS

5.1. Introduction

Medical diagnosis is the process of deducing the disease affecting an individual [139]. This is usually done by clinicians, who analyze the patient's medical record, conduct laboratory tests and physical examination, etc. Accurate diagnosis is essential and quite challenging, as certain diseases have similar symptoms. A good diagnosis should meet some requirements: it should be accurate, communicated, and timely. Misdiagnosis occurs regularly and can be life-threatening; in fact, over 12 million people get misdiagnosed every year in the United States alone [140]. Machine learning (ML) is progressively being applied in medical diagnosis and has achieved significant success so far.

In contrast to the shortfall of clinicians in most countries and expensive manual diagnosis, ML-based diagnosis can significantly improve the healthcare system and reduce misdiagnosis caused by clinicians which can be due to stress, fatigue, and inexperience. Machine learning models can also ensure that patient data are examined in more detail and results obtained quickly [141]. Hence, several researchers and industry experts have developed numerous medical diagnosis models using machine learning [142]. However, some factors are hindering the growth of ML in the medical domain, i.e., the imbalanced nature of medical data and the high cost of labelling data. Imbalanced data is a classification problem in which the number of instances per class is not uniformly distributed. Recently, unsupervised feature learning methods have received massive attention since they do not entirely rely on labelled data [143], and are suitable for training models when the data is imbalanced.

There are various methods used to achieve feature learning, including supervised learning techniques such as dictionary learning and multilayer perceptron (MLP), and unsupervised learning techniques which includes independent component analysis, matrix factorization, clustering, unsupervised dictionary learning, and autoencoders. An autoencoder is a neural network used for unsupervised feature learning. It is composed of input, hidden, and output

UNIVERSITY OF JOHANNESBURG

layers [144]. The basic architecture of a three-layer autoencoder (AE) is shown in Figure 5.1. When given an input data, autoencoders (AEs) are helpful to automatically discover the features that lead to optimal classification [145]. There are diverse forms of autoencoders, including variational and regularized autoencoders. The regularized autoencoders have been mostly used in solving problems where optimal feature learning is needed for subsequent classification, which is the focus of this research. Examples of regularized autoencoders include denoising, contractive, and sparse autoencoders. We aim to implement a sparse autoencoder (SAE) to learn representations more efficiently from raw data in order to ease the classification process and ultimately improve the prediction performance of the classifier.

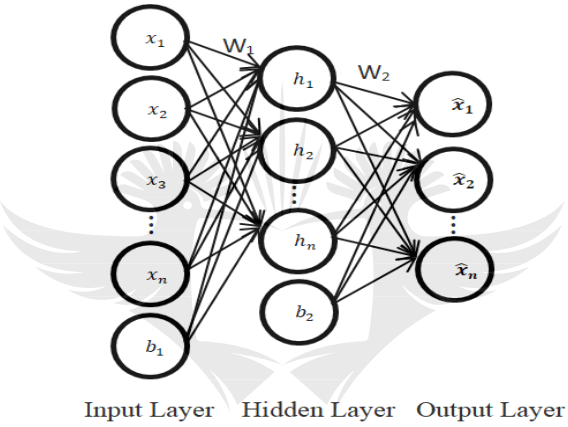


Figure 5.1: The structure of an autoencoder

Usually, the sparsity penalty in the sparse autoencoder network is achieved using either of these two methods: L1 regularization or Kullback-Leibler (KL) divergence. It is noteworthy that the SAE does not regularize the weights of the network; rather, the regularization is imposed on the activations. Consequently, suboptimal performances are obtained with this type of structures where the sparsity makes it challenging for the network to approximate a near-zero cost function [127]. Therefore, in this chapter, we integrate an improved SAE and a softmax classifier with application to medical diagnosis. The SAE imposes regularization on the weights, instead of the activations as in conventional SAE and the softmax classifier is used for performing the classification task.

To demonstrate the effectiveness of the approach, three publicly available medical datasets would be used, i.e., the chronic kidney disease (CKD) dataset [105], cervical cancer risk factors dataset [100], and Framingham heart study dataset [97]. We also aim to use diverse performance

evaluation metrics to assess the performance of the proposed method, compare it with some techniques available in recent literature and other machine learning algorithms such as logistic regression (LR), classification and regression tree (CART), support vector machine (SVM), k-nearest neighbor (KNN), linear discriminant analysis (LDA), and a conventional softmax classifier. The rest of the chapter is structured as follows: Section 5.2 reviews some related works, while Section 5.3 introduces the methodology and provides a detailed background of the methods applied. The results are tabulated and discussed in Section 5.4, while Section 5.5 concludes the chapter.

5.2. Related Works

This section discusses some recent applications of machine learning to medical diagnosis. Glaucoma is a vision condition that develops gradually and can lead to permanent vision loss. This condition destroys the optic nerve, the health of which is essential for good vision and is usually caused by too much pressure inside one or both eyes. There are diverse forms of glaucoma, and they have no warning signs; hence, early detection is difficult yet crucial. Recently, a method was developed for the early detection of glaucoma using a two-layer sparse autoencoder [145]. The SAE was trained using 1426 fundus images to identify salient features from the data and differentiate a normal eye from an affected eye. The structure of the network comprises of two cascaded autoencoders and a softmax layer. The autoencoder network performed unsupervised feature learning, while the softmax was trained in a supervised fashion. The proposed method obtained excellent performance with an F-measure of 0.95.

In another research, a two-stage approach was proposed for the prediction of heart disease using a sparse autoencoder and artificial neural network (ANN) [96]. Unsupervised feature learning was performed with the help of the sparse autoencoder, which was optimized using the adaptive moment estimation (Adam) algorithm, whereas the ANN was used as the classifier. The method achieved an accuracy of 90% on the Framingham heart disease dataset and 98% on the cervical cancer risk factors dataset, which outperformed some ML algorithms. In a similar research, a hybrid technique was proposed for the classification of heart disease where optimal features were selected via the particle swarm optimization (PSO) search technique and k-means clustering [146]. Several supervised learning methods, including decision tree, MLP, and softmax regression, were then utilized for the classification task. The method was tested using a dataset

UNIVERSITY OF JOHANNESBURG

containing 335 cases and 26 attributes, and the experimental results revealed that the hybrid model enhanced the accuracy of the various classifiers, with the softmax regression model obtaining the best performance with 88.4% accuracy.

In [147], an ensemble learning method was developed for the diagnosis of heart disease. The ensemble method was developed via a stacked structure, whereby the base learners were also ensembles. The base learners include gradient boosting, random forest (RF), and extreme gradient boosting (XGBoost). Additionally, feature ranking and selection were conducted using correlation-based feature selection and PSO, respectively. When tested on different heart disease datasets, the proposed method outperformed the conventional ensemble methods. Furthermore, an ensemble learning classifier was recently developed to detect cervical cancer risk [148]. The model comprises of CART, KNN, SVM, and naïve Bayes (NB) as base learners, and the ensemble model achieved an accuracy of 87%.

The application of sparse autoencoders in the medical domain has been widely studied, especially for disease prediction [96]. Furthermore, sparse autoencoders have been utilized for classifying Parkinson's disease (PD). Recently, [149] proposed an approach which involved a feature extraction step using sparse autoencoder, to classify PD efficiently. Prior to the feature extraction, the data was preprocessed, and appropriate input subset selected from the vocal features via an adaptive grey wolf optimization method. After the feature extraction by the SAE, six ML classifiers were then applied to perform the classification task, and the experimental results signaled improved performance compared to other related works.

From the above-related works, we observed that most of the studies have some limitations: firstly, most of the authors utilized a single medical dataset to validate the performance of their models and not many studies experimented on more than two different diseases. By training and testing the model on two or more datasets, appropriate and more reliable conclusions can be drawn, and it can further validate the generalization ability of the ML method. Secondly, some recent research works have implemented sparse autoencoders for feature learning; however, most of these methods achieved sparsity by regularizing the activations [150], which is the norm. However, in this chapter, sparsity is achieved via weight regularization. Also, poor generalization of ML algorithms resulting from imbalanced datasets, which is common in medical data, can be easily addressed using an effective feature learning method such as this.

UNIVERSITY OF JOHANNESBURG

5.3. Methodology

The sparse autoencoder (SAE) is an unsupervised learning method which is used to automatically learn features from unlabeled data. In this type of autoencoder, the training criterion involves a sparsity penalty. Generally, the loss function of an SAE is constructed by penalizing activations within the hidden layers. For any particular sample, the network is encouraged to learn an encoding by activating only a small number of nodes. By introducing sparsity constraints on the network, such as limiting the number of hidden units, the algorithm can learn better relationships from the data [151]. An autoencoder consists of two functions: an encoder and decoder functions. The encoder maps the *d*-dimensional input data to obtain a hidden representation. In contrast, the decoder maps the hidden representation back to a *d*-dimensional vector that is as close as possible to the encoder input [96], [130]. Assuming m denotes the input features, and n represents the neurons of the hidden layer; the encoding and decoding process can be represented with the following equations:

$$a^1 = \begin{bmatrix} a_1^1 \\ \vdots \\ a_n^1 \end{bmatrix} = \begin{bmatrix} w_{1,1}^1 & w_{1,2}^1 & \cdots & w_{1,m}^1 \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1}^1 & w_{n,2}^1 & \cdots & w_{n,m}^1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} + \begin{bmatrix} b_1^1 \\ \vdots \\ b_n^1 \end{bmatrix} \tag{5.1}$$

$$a^2 = \begin{bmatrix} a_1^2 \\ a_2^2 \\ \vdots \\ a_m^2 \end{bmatrix} = \begin{bmatrix} w_{1,1}^2 & \cdots & w_{1,n}^2 \\ \vdots & \ddots & \vdots \\ w_{m,1}^2 & \cdots & w_{m,n}^2 \end{bmatrix} \begin{bmatrix} a_1^1 \\ \vdots \\ a_n^1 \end{bmatrix} + \begin{bmatrix} b_1^2 \\ \vdots \\ b_m^2 \end{bmatrix} \tag{5.2}$$

UNIVERSITY OF JOHANNESBURG

where $w^1 \in R^{n,m}$ and $w^2 \in R^{m,n}$ represents the weight matrices of the hidden layer and output layer, respectively, $b^1 \in R^{n,1}$ and $b^2 \in R^{m,1}$ denotes the bias matrices of the hidden layer and output layer, respectively, the vector $a^1 \in R^{n,1}$ denotes the inputs of the output layer, and the vector $a^2 \in R^{m,1}$ represents the output of the sparse autoencoder, which is fed into the softmax classifier for classification. The mean squared error function E_{MSE} is used as the reconstruction error function between the input x_i and reconstructed input a_i^2. Also, we introduce a regularization function $\Omega_{sparsity}$ to the error function in order to achieve sparsity by penalizing the weights $w^1 \in R^{n,m}$ and $w^2 \in R^{m,n}$. Therefore, the cost function E_{SAE} of the sparse autoencoder can be represented as:

$$E_{SAE} = E_{MSE} + \Omega_{sparsity} \tag{5.3}$$

The mean squared error function and the regularization function can be expressed as:

$$E_{MSE} = \frac{1}{m}\sum_{i=1}^{m}(x_i - a_i^2)^2 \tag{5.4}$$

$$\Omega_{sparsity} = \frac{1}{m}\sum_{i=1}^{m}\left((x_i + 10)\log\frac{x_i+10}{a_i^2+10} + (10 - x_i)\log\frac{10-x_i}{10-a_i^2}\right) \tag{5.5}$$

Once the data has been transmitted from input to output of the sparse autoencoder, the next stage involves evaluating the cost function and fine-tuning the model parameters for optimal performance. Meanwhile, the cost function E_{SAE} does not explicitly relate the weights and bias of the network; hence, it is necessary to define a sensitivity measure to sensitize the changes in E_{SAE} and transmit the changes backwards via the backpropagation learning method [127]. To achieve this, and iteratively optimize the loss function, stochastic gradient descent is employed. The stochastic gradient descent to update the bias and weights of the output layer can be written as:

$$b^2 = b^2 - \eta^2\frac{\partial E_{SAE}}{\partial b^2} \tag{5.6}$$

$$w^2 = w^2 - \eta^2\frac{\partial E_{SAE}}{\partial w^2} \tag{5.7}$$

where η^2 represents the learning rate in relation to the output layer. The derivative of the loss function E_{SAE} measures the sensitivity to change of the function value with respect to a change in its input value. Furthermore, the gradient indicates the extent to which the input parameter needs to change to minimize the loss function. Meanwhile, the gradients are computed using the chain rule. Therefore (5.6) and (5.7) can be rewritten as:

$$b^2 = b^2 - \eta^2\frac{\partial E_{SAE}}{\partial a^2}\times\frac{\partial a^2}{\partial b^2} \tag{5.8}$$

$$w^2 = w^2 - \eta^2\frac{\partial E_{SAE}}{\partial a^2}\times\frac{\partial a^2}{\partial w^2} \tag{5.9}$$

The sensitivity at the output layer of the SAE is represented and defined as $S^2 = \frac{\partial E_{SAE}}{\partial a^2}$. Therefore, (5.8) and (5.9) can be rewritten as:

$$b^2 = b^2 - \eta^2 s^2 \tag{5.10}$$

$$w^2 = w^2 - \eta^2 s^2 (a^1)^T \tag{5.11}$$

where

UNIVERSITY OF JOHANNESBURG

$$S^2 = \begin{bmatrix} s_1^2 \\ s_2^2 \\ \vdots \\ s_m^2 \end{bmatrix} = \begin{bmatrix} \frac{-(x_1+10)}{\log(10)(a_1^2+10)} + \frac{(10-x_1)}{\log(10)(10-a_1^2)} - (x_1 - a_1^2) \\ \frac{-(x_2+10)}{\log(10)(a_2^2+10)} + \frac{(10-x_2)}{\log(10)(10-a_2^2)} - (x_2 - a_2^2) \\ \vdots \\ \frac{-(x_m+10)}{\log(10)(a_m^2+10)} + \frac{(10-x_m)}{\log(10)(10-a_m^2)} - (x_m - a_m^2) \end{bmatrix} \quad (5.12)$$

Using the same method for computing S^2, the sensitivities can be transmitted back to the hidden layer:

$$b^1 = b^1 - \eta^1 s^1 \quad (5.13)$$

$$w^1 = w^1 - \eta^1 s^1 (x)^T \quad (5.14)$$

where η^1 denotes the learning rate with respect to the hidden layer, whereas s^1 is defined as:

$$s^1 = \begin{bmatrix} s_1^1 \\ \vdots \\ s_n^1 \end{bmatrix} = \begin{bmatrix} s_1^2 w_{1,1}^2 + s_2^2 w_{2,1}^2 + \cdots + s_m^2 w_{m,1}^2 \\ \vdots \\ s_1^2 w_{1,n}^2 + s_2^2 w_{2,n}^2 + \cdots + s_m^2 w_{m,n}^2 \end{bmatrix} \quad (5.15)$$

Furthermore, the softmax classifier is employed for the classification task. The learned features from the proposed SAE are used to train the classifier. Though, softmax regression, otherwise called multinomial logistic regression (MLR), is a generalization of logistic regression that can be utilized for multi-class classification [152]. However, the softmax classifier has been applied for several binary classification tasks and obtained excellent performance [153]. The softmax function provides a method to interpret the outputs as probabilities and is expressed as:

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}} \ (i = 1,2, \dots, N) \quad (5.16)$$

where $x_1, x_2, \dots, x_N$ represent the input values and the output $f(x_i)$ is the probability that the sample belongs to the *i*-th label [65]. For N input samples, the error at the softmax layer is measured using the cross-entropy loss function:

$$L(w) = \frac{1}{N}\sum_{n=1}^{N} H(p_n, q_n) = -\frac{1}{N}\sum_{n=1}^{N}[y_n \log \hat{y}_n + (1 - y_n)\log(1 - \hat{y}_n)] \quad (5.17)$$

UNIVERSITY OF JOHANNESBURG

where the true probability p_n is the actual label and q_n is the predicted value. $H(p_n, q_n)$ is a measure of the dissimilarity between p_n and q_n. Furthermore, neural networks can easily get stuck in local minima, whereby the algorithm assumes it has reached the global minima, thereby resulting in non-optimal performance. To prevent the local minima problem and further enhance the classifier performance, the mini-batch gradient descent with momentum is applied to optimize the cross-entropy loss of the softmax classifier. This optimization algorithm splits the training data into small batches which are then used to compute the model error and update the model parameters [154]. The momentum [155] ensures better convergence is obtained.

The flowchart to visualize the proposed methodology is shown in Figure 5.2. The initial dataset is preprocessed; then divided into training and testing sets. The training set is utilized for training the sparse autoencoder in an unsupervised manner. Meanwhile, the testing set is transformed and inputted into the trained model to obtain the low-dimensional representation dataset. The low-dimensional training set is used to train the softmax classifier, and its performance is tested using the low-dimensional test set. Hence, there is no possible data leakage since the classifier sees only the low-dimensional training set.

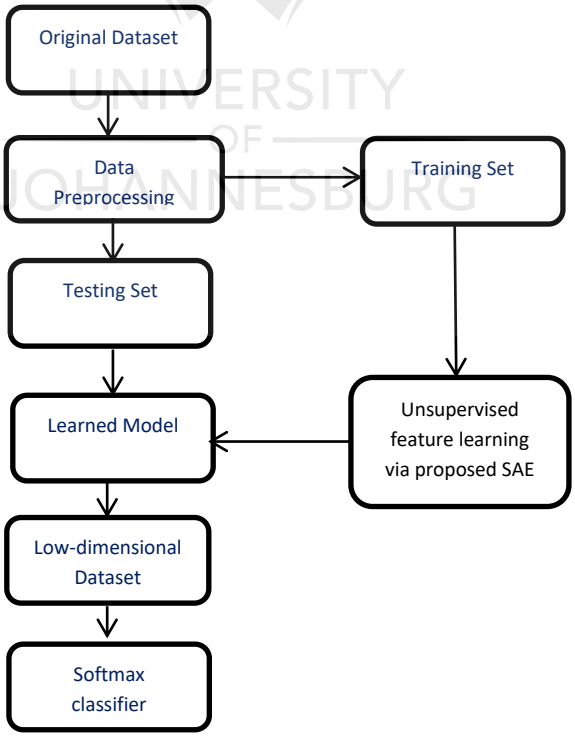


Figure 5.2: Flowchart of the methodology

5.4. Results and Discussion

The proposed method is applied for the prediction of three diseases in order to show its performance in diverse medical diagnosis situations. The training parameters of the SAE include: $\eta^1 = 0.01, \eta^2 = 0.1, n = 25$, and number of epochs = 200. The hyperparameters of the softmax classifier include: learning rate = 0.01, number of samples in mini-batches = 32, momentum value = 0.9 and number of epochs = 200. These parameters were obtained from the literature [96], [154], as they have led to optimal performance in diverse neural network applications.

To demonstrate the efficacy of the proposed method, it is benchmarked with other algorithms, such as LR, CART, SVM, KNN, LDA, and conventional softmax regression. In order to show the improved performance of the proposed method, no parameter tuning was performed on these algorithms; hence, their default parameter values in scikit-learn were used, which are adequate for most machine learning problems. The K-fold cross-validation technique was used to evaluate all the models. Tables 5.1, 5.2, and 5.3 show the experimental results when the proposed method is tested on the Framingham heart study, cervical cancer risk factors, and CKD datasets, respectively. While, Figures 5.3, 5.4, and 5.5 show the ROC curves comparing the performance of the conventional softmax classifier and the proposed approach for the various disease prediction models. The ROC curve illustrates the diagnostic ability of binary classifiers, and it is obtained by plotting the true positive rate (TPR) against the false positive rate (FPR).

UNIVERSITY OF JOHANNESBURG

Table 5.1: Performance of the proposed method and other classifiers on the Framingham dataset

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
LR	83	84	86	84
CART	75	74	75	74
SVM	82	78	82	80
KNN	81	75	81	77
LDA	83	81	83	82
Softmax classifier	86	84	88	86
Proposed SAE+Softmax	91	93	90	92

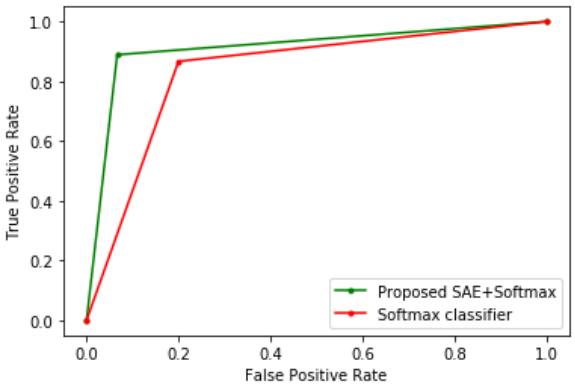
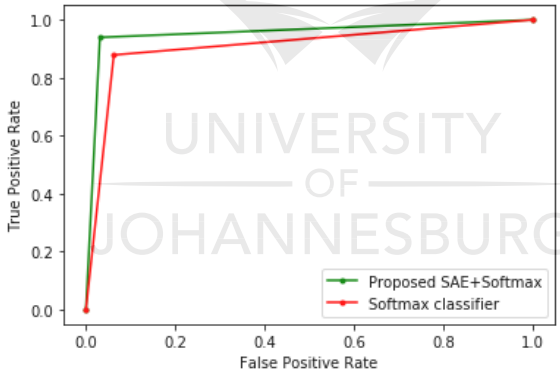


Figure 5.3: ROC curve of the heart disease model

Table 5.2: Performance of the proposed method and other classifiers on the cervical cancer dataset

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
LR	94	96	91	93
CART	90	93	96	94
SVM	94	90	93	91
KNN	93	98	95	96
LDA	95	93	91	92
Softmax classifier	94	97	91	94
Proposed SAE+Softmax	97	98	95	97



UNIVERSITY OF JOHANNESBURG

Figure 5.4: ROC curve of the cervical cancer model

Table 5.3: Performance of the proposed method and other classifiers on the CKD dataset

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
LR	98	93	97	95
CART	95	97	95	96
SVM	96	94	96	95
KNN	94	93	89	91
LDA	96	97	93	95
Softmax classifier	96	95	97	96
Proposed SAE+Softmax	98	97	97	97

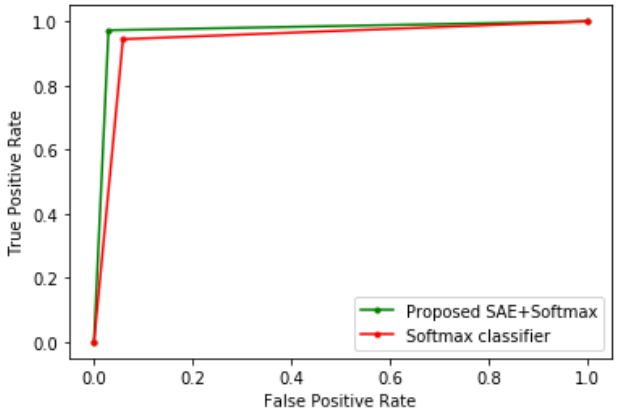


Figure 5.5: ROC curve of the CKD model

From the experimental results, it can be seen that the sparse autoencoder improves the performance of the softmax classifier, which is further validated by the ROC curves of the various models. The proposed method also performed better than the other machine learning algorithms. Furthermore, the misclassifications obtained by the model in the various disease predictions are also considered. For the prediction of heart disease, the proposed method recorded FPR of 7% and false-negative rate (FNR) of 10%. In addition, the model specificity, which is the true negative rate (TNR) is 93%, and the TPR is 90%. For the cervical cancer dataset, the following were obtained: FPR = 3%, FNR = 5%, TNR = 97%, and TPR = 95%. For the CKD prediction: FPR = 0, FNR = 3%, TNR = 100%, and TPR = 97%.

Additionally, to further validate the performance of the proposed method, we compare it with some models for heart disease prediction available in recent literature, including a feature selection method using PSO and softmax regression [146], a two-tier ensemble method with PSO based feature selection [147], an ensemble classifier comprising of the following base learners: NB, Bayes Net (BN), RF, and MLP [92], a hybrid method of NB and LR [156], and a hybrid RF with a linear model (HRFLM) [157]. The other techniques include a combination of LR and Lasso regression [93], an intelligent heart disease detection method based on NB and advanced encryption standard (AES) [158], a combination of ANN and Fuzzy analytic hierarchy method (Fuzzy-AHP) [159], and a sparse autoencoder feature learning method combined ANN classifier [12]. This comparison is tabulated in Table 5.4. Meanwhile, in order to give a fair comparison,

UNIVERSITY OF JOHANNESBURG

only the accuracies of the various techniques were considered because some authors did not report the values for other performance metrics.

Table 5.4: Comparison of the proposed method with recent literature that used the heart disease dataset

Algorithm	Method	Accuracy (%)
Verma et al. [146]	PSO and Softmax regression	88.4
Tama et al. [147]	Ensemble and PSO	85.71
Latha and Jeeva [92]	An Ensemble of NB, BN, RF, and MLP	85.48
Amin et al. [156]	A hybrid NB and LR	87.4
Mohan et al. [157]	HRFLM	88.4
Haq et al. [93]	LASSO-LR Model	89
Repaka et al. [158]	NB-AES	89.77
Samuel et al. [159]	ANN-Fuzzy-AHP	91
Mienye et al. [96]	SAE+ANN	90
Our approach	Improved SAE+Softmax	91

In Table 5.5, we compare the proposed approach with some recent scholarly works that used the cervical cancer dataset, including principal component analysis (PCA) based SVM [101], a research work where the dataset was preprocessed and classified using numerous algorithms, in which LR and SVM had the best accuracy [160], and a C5.0 decision tree [161]. The other methods include a multistage classification process which combined isolation forest (iForest), the synthetic minority over-sampling technique (SMOTE), and RF [162], a sparse autoencoder feature learning method combined with an ANN classifier [12], and a feature selection method combined with C5.0 and RF [163].

UNIVERSITY OF JOHANNESBURG

Table 5.5: Comparison of the proposed method with recent literature that used the cervical cancer dataset

Algorithm	Method	Accuracy (%)
Wu and Zhou [101]	SVM-PCA	94.03
Abdullah et al. [160]	SVM	93.4884
	LR	93.4884
Chang et al. [161]	C5.0	96
Ijaz et al. [162]	iForest+SMOTE+RF	98.925
Mienye et al. [96]	SAE+ANN	98
Nithya and Ilango [163]	C5.0	97
	RF	96.9
Our approach	Improved SAE+Softmax	97

In Table 5.6, we compare the proposed method with other recent CKD prediction research works, including an optimized XGBoost method [104], a probabilistic neural network (PNN)

[164], and a method using adaptive boosting (AdaBoost) [165]. The other research works include a hybrid classifier of NB and decision tree (NBTree) [166], XGBoost [167], and a 7-7-1 MLP neural network [168].

Table 5.6: Comparison of the proposed method with recent literature that used the CKD dataset

Algorithm	Method	Accuracy (%)
Ogunleye and Qing-Guo [104]	Optimized XGBoost	100
Rady and Anwar [164]	PNN	96.7
Gupta et al. [165]	AdaBoost	88.66
Khan et al. [166]	NBTree	98.75
Raju et al. [167]	XGBoost	99.29
Aljaaf et al. [168]	MLP	98.1
Our approach	Improved SAE+Softmax	98

The proposed sparse autoencoder with softmax regression obtained comparable performance with the state-of-the-art methods in various disease predictions from the tabulated comparisons. However, it was observed that in Table 5.5, a few methods achieved slightly better performance than the proposed approach; for example, in [162], the performance can be attributed to the data preprocessing, where the authors performed outlier detection and oversampling before classification. And in [96][166], the batch normalization technique was utilized to enhance the model's performance. Also, from Table 5.6, [104], [166], [167] obtained better results as their approaches combined multiple classification algorithms.

UNIVERSITY OF JOHANNESBURG

The experimental results also show an improved performance obtained due to efficient feature representation by the sparse autoencoder. This further demonstrates the importance of training classifiers with relevant data since it can significantly affect the performance of the prediction model. This research also showed that excellent classification performance could be obtained not only by performing hyperparameter tuning of algorithms but also by employing appropriate feature learning techniques. The proposed models could also be used for multi-class classification since the softmax regression works under the assumption that the classes are mutually exclusive.

5.5. Conclusion

In this chapter, we developed an approach for improved prediction of diseases based on an enhanced sparse autoencoder and softmax regression. Usually, autoencoders achieve sparsity by penalizing the activations within the hidden layers, but in the proposed method, the weights were

penalized instead. This is necessary because penalizing the activations makes approximating a near-zero loss function challenging for the network. The proposed method was tested on three different diseases, including heart disease, cervical cancer, and chronic kidney disease, and it achieved accuracies of 91%, 97%, and 98% respectively, which outperformed conventional softmax regression and other algorithms. By experimenting with different datasets, we aimed to demonstrate the effectiveness of the method in diverse conditions. We also conducted a comparative study with some prediction models available in recent literature, and the proposed approach obtained comparable performance in terms of accuracy. Thus, it can be concluded that the proposed approach is a promising method for the detection of diseases and can be further developed into a clinical decision support system to assist health professionals. Meanwhile, future research will apply the method studied in this chapter for the prediction of more diseases, and also employ other performance metrics such as training time, classification time, computational speed, and other metrics, which could be beneficial for the performance evaluation of the model. Future works can also employ appropriate data preprocessing technique and combine the proposed feature learning method with an ensemble classifier to further enhance the classification performance. Lastly, future works can also utilize the proposed method for medical image classification since similar sparse autoencoders have been employed for diverse image recognition tasks.

UNIVERSITY
OF
JOHANNESBURG

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1. Conclusion

The dissertation is concluded in this chapter, and a few future research directions are also presented. Machine learning has proven to be a vital tool in several domains and its application for credit card default prediction, and medical diagnosis has progressed rapidly in recent times, with several research works available in the literature. This dissertation discussed several applications of machine learning for credit risk prediction and medical diagnosis, and also developed two enhanced methods. Chapter 1 presented a detailed background on ML and the problem of imbalanced data, and the objectives of the dissertation were also discussed. The outcomes of this dissertation are outlined below, which corresponds to the various objectives of the dissertation.

- Chapter 2 presented an extensive survey of some recent machine learning research works with application to the prediction of credit risk and medical diagnosis. Sections 4.2 and 5.2 also discussed some previous works relevant to Chapters 4 and 5, respectively. Some ML algorithms applied in the course of this research were also discussed in Chapter 2 to lay a solid foundation for the dissertation, and their mathematical representations were also discussed in detail. Chapter 3 provided the research methodology and discussed the datasets and performance metrics used in the research.
- A method to enhance the prediction of credit card default was presented in Chapter 4 using a stacked sparse autoencoder, which learned the best representation of the input data in order to train the classifiers with the most relevant data that will result in improved performance. Batch normalization was introduced to the autoencoder network to address the problem of internal covariate shift which usually occurs in deep neural networks. The stacked sparse autoencoder combined with linear discriminant analysis obtained the best performance with an accuracy of 90%, precision of 91%, recall of 90%, and F1 score of 90%.

UNIVERSITY OF JOHANNESBURG

- In Chapter 5, a method was developed for the prediction of heart disease, cervical cancer, and chronic kidney disease. The approach integrates an enhanced sparse autoencoder with softmax regression. The improved method obtained better performance compared to other algorithms and some recently proposed scholarly works: for heart disease (accuracy = 91%, precision = 93%, recall = 90%, and F1 score = 92%), for cervical cancer (accuracy = 97%, precision = 98%, recall = 95%, and F1 score = 97%), and for chronic kidney disease (accuracy = 98%, precision = 97%, recall = 97%, and F1 score = 97%).

The rationale behind using these feature learning methods to improve the performance of the classifiers was based on the fact that the data used to train machine learning algorithms impacts on the final performance, and this was confirmed as the proposed methods obtained better performance than the conventional machine learning algorithms. The proposed methods also showed comparable performance with several recently developed methods available in the literature.

6.2. Future works

Future research works would focus on how best to deploy unsupervised feature learning techniques in real-world applications where the data is imbalanced. This is important because supervised learning requires domain knowledge for feature engineering, which is time-consuming and expensive. Therefore, future research could integrate feature learning techniques in diverse medical diagnosis decision support systems which are mainly supervised. Furthermore, future research works can also consider other performance metrics suited for imbalanced classification such as geographic mean or G-mean, Fbeta-measure, balanced accuracy, Kappa, etc. Also, future research could be done more efficiently with a considerable amount of resources, including time, processing power and memory: this can be achieved through the use of GPUs. Therefore, GPUs could be beneficial for future research.

UNIVERSITY OF JOHANNESBURG

REFERENCES

[1] M. Raj and R. Seamans, "Primer on artificial intelligence and robotics," *Journal of Organization Design*, vol. 8, no. 1, p. 11, May 2019, doi: 10.1186/s41469-019-0050-0.

[2] S. Raschka, J. Patterson, and C. Nolet, "Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence," *Information*, vol. 11, no. 4, Art. no. 4, Apr. 2020, doi: 10.3390/info11040193.

[3] G. Shobha and S. Rangaswamy, "Chapter 8 - Machine Learning," in *Handbook of Statistics*, vol. 38, V. N. Gudivada and C. R. Rao, Eds. Elsevier, 2018, pp. 197–228.

[4] T. Pan, J. Zhao, W. Wu, and J. Yang, "Learning imbalanced datasets based on SMOTE and Gaussian distribution," *Information Sciences*, vol. 512, pp. 1214–1233, Feb. 2020, doi: 10.1016/j.ins.2019.10.048.

[5] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Information Sciences*, vol. 513, pp. 429–441, Mar. 2020, doi: 10.1016/j.ins.2019.11.004.

[6] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/j.patcog.2019.02.023.

[7] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, Oct. 2018, doi: 10.1016/j.neunet.2018.07.011.

[8] K. K. Bejjanki, J. Gyani, and N. Gugulothu, "Class Imbalance Reduction (CIR): A Novel Approach to Software Defect Prediction in the Presence of Class Imbalance," *Symmetry*, vol. 12, no. 3, Art. no. 3, Mar. 2020, doi: 10.3390/sym12030407.

[9] C. Bellinger, S. Sharma, N. Japkowicz, and O. R. Zaïane, "Framework for extreme imbalance classification: SWIM—sampling with the majority class," *Knowl Inf Syst*, vol. 62, no. 3, pp. 841–866, Mar. 2020, doi: 10.1007/s10115-019-01380-z.

[10] J. Hernandez, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "An Empirical Study of Oversampling and Undersampling for Instance Selection Methods on Imbalance Datasets," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, Berlin, Heidelberg, 2013, pp. 262–269, doi: 10.1007/978-3-642-41822-8_33.

[11] H. Patel, D. Singh Rajput, G. Thippa Reddy, C. Iwendi, A. Kashif Bashir, and O. Jo, "A review on classification of imbalanced data for wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 16, no. 4, p. 1550147720916404, Apr. 2020, doi: 10.1177/1550147720916404.

[12] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, Jul. 2012, doi: 10.1109/TSMCC.2011.2161285.

[13] Q. Wang, Z. Luo, J. Huang, Y. Feng, and Z. Liu, "A Novel Ensemble Method for Imbalanced Data Learning: Bagging of Extrapolation-SMOTE SVM," *Computational Intelligence and Neuroscience*, Jan. 30, 2017. https://www.hindawi.com/journals/cin/2017/1827016/ (accessed Aug. 09, 2020).

[14] W. Feng, W. Huang, and J. Ren, "Class Imbalance Ensemble Learning Based on the Margin Theory," *Applied Sciences*, vol. 8, no. 5, Art. no. 5, May 2018, doi: 10.3390/app8050815.

[15] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *Journal of Big Data*, vol. 5, no. 1, p. 42, Nov. 2018, doi: 10.1186/s40537-018-0151-6.

[16] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J Big Data*, vol. 6, no. 1, p. 27, Mar. 2019, doi: 10.1186/s40537-019-0192-5.

[17] S. A. Ebiaredoh-Mienye, E. Esenogho, and T. G. Swart, "Integrating Enhanced Sparse Autoencoder-Based Artificial Neural Network Technique and Softmax Regression for Medical Diagnosis," *Electronics*, vol. 9, no. 11, Art. no. 11, Nov. 2020, doi: 10.3390/electronics9111963.
[18] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013, doi: 10.1109/TPAMI.2013.50.
[19] C. T. Sari and C. Gunduz-Demir, "Unsupervised Feature Extraction via Deep Learning for Histopathological Classification of Colon Tissue Images," *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1139–1149, May 2019, doi: 10.1109/TMI.2018.2879369.
[20] S. Hussein, P. Kandel, C. W. Bolan, M. B. Wallace, and U. Bagci, "Lung and Pancreatic Tumor Characterization in the Deep Learning Era: Novel Supervised and Unsupervised Learning Approaches," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1777–1787, Aug. 2019, doi: 10.1109/TMI.2019.2894349.
[21] S. Yang, Y. Zhang, Y. Zhu, P. Li, and X. Hu, "Representation learning via serial autoencoders for domain adaptation," *Neurocomputing*, vol. 351, pp. 1–9, Jul. 2019, doi: 10.1016/j.neucom.2019.03.056.
[22] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota, "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction," *Information Sciences*, vol. 428, pp. 49–61, Feb. 2018, doi: 10.1016/j.ins.2017.10.044.
[23] T. Marwala and B. Xing, "Blockchain and Artificial Intelligence," *arXiv:1802.04451 [cs]*, Oct. 2018, Accessed: Aug. 04, 2020. [Online]. Available: http://arxiv.org/abs/1802.04451.
[24] R. Cioffi, M. Travaglioni, G. Piscitelli, A. Petrillo, and F. De Felice, "Artificial Intelligence and Machine Learning Applications in Smart Production: Progress, Trends, and Directions," *Sustainability*, vol. 12, no. 2, Art. no. 2, Jan. 2020, doi: 10.3390/su12020492.
[25] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," *BMC Medical Research Methodology*, vol. 19, no. 1, p. 64, Mar. 2019, doi: 10.1186/s12874-019-0681-4.
[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[27] E. Aïmeur, G. Brassard, and S. Gambs, "Quantum speed-up for unsupervised learning," *Mach Learn*, vol. 90, no. 2, pp. 261–287, Feb. 2013, doi: 10.1007/s10994-012-5316-5.
[28] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach Learn*, vol. 109, no. 2, pp. 373–440, Feb. 2020, doi: 10.1007/s10994-019-05855-6.
[29] F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection," *Int. j. inf. tecnol.*, Feb. 2020, doi: 10.1007/s41870-020-00430-y.
[30] O. Adepoju, J. Wosowei, S. lawte, and H. Jaiman, "Comparative Evaluation of Credit Card Fraud Detection Using Machine Learning Techniques," in *2019 Global Conference for Advancement in Technology (GCAT)*, Oct. 2019, pp. 1–6, doi: 10.1109/GCAT47503.2019.8978372.
[31] A. Bindal and S. Chaurasia, "Predictive Risk Analysis For Loan Repayment of Credit Card Clients," in *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, May 2018, pp. 2508–2513, doi: 10.1109/RTEICT42901.2018.9012366.
[32] S. Tortajada, M. Robles, and J. M. García-Gómez, "Incremental Logistic Regression for Customizing Automatic Diagnostic Models," in *Data Mining in Clinical Medicine*, C. Fernández-Llatas and J. M. García-Gómez, Eds. New York, NY: Springer, 2015, pp. 57–78.
[33] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatics in Medicine Unlocked*, vol. 17, p. 100179, Jan. 2019, doi: 10.1016/j.imu.2019.100179.

[34] V. J. Ribas, A. Vellido, J. C. Ruiz-Rodríguez, and J. Rello, "Severe sepsis mortality prediction with logistic regression over latent factors," *Expert Systems with Applications*, vol. 39, no. 2, pp. 1937–1943, Feb. 2012, doi: 10.1016/j.eswa.2011.08.054.

[35] A. Keeley, P. Hine, and E. Nsutebu, "The recognition and management of sepsis and septic shock: a guide for non-intensivists," *Postgraduate Medical Journal*, vol. 93, no. 1104, pp. 626–634, Oct. 2017, doi: 10.1136/postgradmedj-2016-134519.

[36] K. Thompson, B. Venkatesh, and S. Finfer, "Sepsis and septic shock: current approaches to management," *Internal Medicine Journal*, vol. 49, no. 2, pp. 160–170, 2019, doi: 10.1111/imj.14199.

[37] M. C. Aniceto, F. Barboza, and H. Kimura, "Machine learning predictivity applied to consumer creditworthiness," *Future Business Journal*, vol. 6, no. 1, p. 37, Nov. 2020, doi: 10.1186/s43093-020-00041-w.

[38] A. Subasi and S. Cankurt, "Prediction of default payment of credit card clients using Data Mining Techniques," in *2019 International Engineering Conference (IEC)*, Jun. 2019, pp. 115–120, doi: 10.1109/IEC47844.2019.8950597.

[39] T. M. Alam *et al.*, "An Investigation of Credit Card Default Prediction in the Imbalanced Datasets," *IEEE Access*, pp. 1–1, 2020, doi: 10.1109/ACCESS.2020.3033784.

[40] L. Tanner *et al.*, "Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness," *PLoS Negl Trop Dis*, vol. 2, no. 3, Mar. 2008, doi: 10.1371/journal.pntd.0000196.

[41] A. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," *Neural Comput & Applic*, vol. 23, no. 7, pp. 2387–2403, Dec. 2013, doi: 10.1007/s00521-012-1196-7.

[42] D. Obare and M. Muraya, "Comparison of Accuracy of Support Vector Machine Model and Logistic Regression Model in Predicting Individual Loan Defaults," *American Journal of Applied Mathematics and Statistics*, vol. 6, no. 6, Art. no. 6, Dec. 2018, doi: 10.12691/ajams-6-6-8.

[43] F. E. Moula, C. Guotai, and M. Z. Abedin, "Credit default prediction modeling: an application of support vector machine," *Risk Manag*, vol. 19, no. 2, pp. 158–187, May 2017, doi: 10.1057/s41283-017-0016-x.

[44] E. Gürbüz and E. Kılıç, "A new adaptive support vector machine for diagnosis of diseases," *Expert Sys: J. Knowl. Eng.*, vol. 31, no. 5, pp. 389–397, Nov. 2014, doi: 10.1111/exsy.12051.

[45] L. Hussain *et al.*, "Detecting Congestive Heart Failure by Extracting Multimodal Features and Employing Machine Learning Techniques," *BioMed Research International*, Feb. 18, 2020. https://www.hindawi.com/journals/bmri/2020/4281243/ (accessed Aug. 08, 2020).

[46] M. A. Mukid, T. Widiharih, A. Rusgiyono, and A. Prahutama, "Credit scoring analysis using weighted k nearest neighbor," *Journal of Physics: Conference Series*, vol. 1025, p. 012114, May 2018, doi: 10.1088/1742-6596/1025/1/012114.

[47] N. Sariannidis, S. Papadakis, A. Garefalakis, C. Lemonakis, and T. Kyriaki-Argyro, "Default avoidance on credit card portfolios using accounting, demographical and exploratory factors: decision making based on machine learning (ML) techniques," *Ann Oper Res*, vol. 294, no. 1, pp. 715–739, Nov. 2020, doi: 10.1007/s10479-019-03188-0.

[48] J. E. Dalen, J. S. Alpert, R. J. Goldberg, and R. S. Weinstein, "The Epidemic of the 20th Century: Coronary Heart Disease," *The American Journal of Medicine*, vol. 127, no. 9, pp. 807–812, Sep. 2014, doi: 10.1016/j.amjmed.2014.04.015.

[49] Qin Yanwen *et al.*, "Mitochondrial tRNA Variants in Chinese Subjects With Coronary Heart Disease," *Journal of the American Heart Association*, vol. 3, no. 1, p. e000437, 2014, doi: 10.1161/JAHA.113.000437.

[50] M. A. jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm," *Procedia Technology*, vol. 10, pp. 85–94, Jan. 2013, doi: 10.1016/j.protcy.2013.12.340.

[51] C. Sowmiya and P. Sumitra, "A hybrid approach for mortality prediction for heart patients using ACO-HKNN," *J Ambient Intell Human Comput*, May 2020, doi: 10.1007/s12652-020-02027-6.
[52] R. Garcia-Carretero, L. Vigil-Medina, I. Mora-Jimenez, C. Soguero-Ruiz, O. Barquero-Perez, and J. Ramos-Lopez, "Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population," *Med Biol Eng Comput*, vol. 58, no. 5, pp. 991–1002, May 2020, doi: 10.1007/s11517-020-02132-w.
[53] A. Krichene, "Using a naive Bayesian classifier methodology for loan risk assessment:Evidence from a Tunisian commercial bank," *Journal of Economics, Finance and Administrative Science*, vol. 22, no. 42, pp. 3–24, Jan. 2017, doi: 10.1108/JEFAS-02-2017-0039.
[54] H. Kamel, D. Abdulah, and J. M. Al-Tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," in *2019 International Engineering Conference (IEC)*, Jun. 2019, pp. 165–170, doi: 10.1109/IEC47844.2019.8950650.
[55] M. A. Arasi, E.-S. M. El-Horbaty, and E.-S. A. E.-D. El-Dahshan, "Classification of Dermoscopy Images Using Naïve Bayesian and Decision Tree Techniques," in *2018 1st Annual International Conference on Information and Sciences (AiCIS)*, Nov. 2018, pp. 7–12, doi: 10.1109/AiCIS.2018.00015.
[56] M. A. Jabbar and S. Samreen, "Heart disease prediction system based on hidden naïve bayes classifier," in *2016 International Conference on Circuits, Controls, Communications and Computing (I4C)*, Oct. 2016, pp. 1–5, doi: 10.1109/CIMCA.2016.8053261.
[57] F. Qian, C. Gong, L. Liu, L. Sha, and M. Zhang, "Topic medical concept embedding: Multi-sense representation learning for medical concept," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2017, pp. 404–409, doi: 10.1109/BIBM.2017.8217683.
[58] H.-Y. Yang and L. H. Staib, "Dual Adversarial Autoencoder for Dermoscopic image Generative Modeling," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Apr. 2019, pp. 1247–1250, doi: 10.1109/ISBI.2019.8759293.
[59] X. Li, M. Radulovic, K. Kanjer, and K. N. Plataniotis, "Discriminative Pattern Mining for Breast Cancer Histopathology Image Classification via Fully Convolutional Autoencoder," *IEEE Access*, vol. 7, pp. 36433–36445, 2019, doi: 10.1109/ACCESS.2019.2904245.
[60] Q. Zhou, B. Yong, Q. Lv, J. Shen, and X. Wang, "Deep Autoencoder for Mass Spectrometry Feature Learning and Cancer Detection," *IEEE Access*, vol. 8, pp. 45156–45166, 2020, doi: 10.1109/ACCESS.2020.2977680.
[61] S. Li, H. Lei, F. Zhou, J. Gardezi, and B. Lei, "Longitudinal and Multi-modal Data Learning for Parkinson's Disease Diagnosis via Stacked Sparse Auto-encoder," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Apr. 2019, pp. 384–387, doi: 10.1109/ISBI.2019.8759385.
[62] Y. Nasser, M. E. Hassouni, A. Brahim, H. Toumi, E. Lespessailles, and R. Jennane, "Diagnosis of osteoporosis disease from bone X-ray images with stacked sparse autoencoder and SVM classifier," in *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, May 2017, pp. 1–5, doi: 10.1109/ATSIP.2017.8075537.
[63] U. Kose and O. Deperlioglu, "Electro-Search Algorithm and Autoencoder Based Recurrent Neural Network for Practical Medical Diagnosis," in *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Oct. 2019, pp. 1–6, doi: 10.1109/ASYU48272.2019.8946427.
[64] W. Wei *et al.*, "Predicting Lymph Node Metastasis of Lung Cancer Using Stacked Sparse Autoencoder," in *2018 14th IEEE International Conference on Signal Processing (ICSP)*, Aug. 2018, pp. 558–561, doi: 10.1109/ICSP.2018.8652453.
[65] M. Wang, S. Lu, D. Zhu, J. Lin, and Z. Wang, "A High-Speed and Low-Complexity Architecture for Softmax Function in Deep Learning," in *2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, Oct. 2018, pp. 223–226, doi: 10.1109/APCCAS.2018.8605654.

UNIVERSITY OF JOHANNESBURG

[66] A. Topîrceanu and G. Grosseck, "Decision tree learning used for the classification of student archetypes in online courses," *Procedia Computer Science*, vol. 112, pp. 51–60, Jan. 2017, doi: 10.1016/j.procs.2017.08.021.
[67] L. Rokach and O. Maimon, "Decision Trees," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Boston, MA: Springer US, 2005, pp. 165–192.
[68] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and Regression Trees," *Wadsworth & Brooks, Monterey*, 1983. /paper/Classification-and-Regression-Trees-Breiman-Friedman/8017699564136f93af21575810d557dba1ee6fc6 (accessed Aug. 05, 2020).
[69] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.
[70] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Elsevier, 2014.
[71] C. Kingsford and S. L. Salzberg, "What are decision trees?," *Nat Biotechnol*, vol. 26, no. 9, pp. 1011–1013, Sep. 2008, doi: 10.1038/nbt0908-1011.
[72] M. Krzywinski and N. Altman, "Classification and regression trees," *Nature Methods*, vol. 14, no. 8, Art. no. 8, Aug. 2017, doi: 10.1038/nmeth.4370.
[73] T. Marwala, *Artificial Intelligence Techniques for Rational Decision Making*. Springer, 2014.
[74] M. Kafai and K. Eshghi, "CROification: Accurate Kernel Classification with the Efficiency of Sparse Linear SVM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 34–48, Jan. 2019, doi: 10.1109/TPAMI.2017.2785313.
[75] S. Hongmao, "Chapter 5 - Quantitative Structure–Activity Relationships: Promise, Validations, and Pitfalls," in *A Practical Guide to Rational Drug Design*, S. Hongmao, Ed. Woodhead Publishing, 2016, pp. 163–192.
[76] H. Saadatfar, S. Khosravi, J. H. Joloudari, A. Mosavi, and S. Shamshirband, "A New K-Nearest Neighbors Classifier for Big Data Based on Efficient Data Pruning," *Mathematics*, vol. 8, no. 2, Art. no. 2, Feb. 2020, doi: 10.3390/math8020286.
[77] I. M. Galván, J. M. Valls, N. Lecomte, and P. Isasi, "A Lazy Approach for Machine Learning Algorithms," in *Artificial Intelligence Applications and Innovations III*, vol. 296, Iliadis, Maglogiann, Tsoumakasis, Vlahavas, and Bramer, Eds. Boston, MA: Springer US, 2009, pp. 517–522.
[78] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of Translational Medicine*, vol. 4, no. 11, Art. no. 11, Apr. 2016, doi: 10.21037/atm.2016.03.37.
[79] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Systems*, vol. 192, p. 105361, Mar. 2020, doi: 10.1016/j.knosys.2019.105361.
[80] Y. Huang and L. Li, "Naive Bayes classification algorithm based on small sample set," in *2011 IEEE International Conference on Cloud Computing and Intelligence Systems*, Sep. 2011, pp. 34–39, doi: 10.1109/CCIS.2011.6045027.
[81] G. Zhong, L.-N. Wang, X. Ling, and J. Dong, "An overview on data representation learning: From traditional feature learning to recent deep learning," *The Journal of Finance and Data Science*, vol. 2, no. 4, pp. 265–278, Dec. 2016, doi: 10.1016/j.jfds.2017.05.001.
[82] W. H. Lopez Pinaya, S. Vieira, R. Garcia-Dias, and A. Mechelli, "Chapter 11 - Autoencoders," in *Machine Learning*, A. Mechelli and S. Vieira, Eds. Academic Press, 2020, pp. 193–208.
[83] M. Hamadache, J. H. Jung, J. Park, and B. D. Youn, "A comprehensive review of artificial intelligence-based approaches for rolling element bearing PHM: shallow and deep learning," *JMST Advances*, vol. 1, no. 1–2, pp. 125–151, Jun. 2019, doi: 10.1007/s42791-019-0016-y.
[84] S. Berardo, E. Favero, and N. Neto, "Active Learning with Clustering and Unsupervised Feature Learning," in *Advances in Artificial Intelligence*, Cham, 2015, pp. 281–290, doi: 10.1007/978-3-319-18356-5_25.
[85] X. Chen, Y. Xu, S. Yan, D. W. K. Wong, T. Y. Wong, and J. Liu, "Automatic Feature Learning for Glaucoma Detection Based on Deep Learning," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Cham, 2015, pp. 669–677, doi: 10.1007/978-3-319-24574-4_80.

[86] X. Gao, S. Lin, and T. Y. Wong, "Automatic Feature Learning to Grade Nuclear Cataracts Based on Deep Learning," in *Computer Vision -- ACCV 2014*, Cham, 2015, pp. 632–642, doi: 10.1007/978-3-319-16808-1_42.

[87] X. Xie, X. Jiang, W. Wang, B. Wang, T. Wan, and H. Yang, "An Intrusion Detection Method Based on Hierarchical Feature Learning and Its Application," in *Cyberspace Safety and Security*, Cham, 2019, pp. 13–20, doi: 10.1007/978-3-030-37337-5_2.

[88] H. Zhao, Z. Lai, H. Leung, and X. Zhang, "A Gentle Introduction to Feature Learning," in *Feature Learning and Understanding: Algorithms and Applications*, H. Zhao, Z. Lai, H. Leung, and X. Zhang, Eds. Cham: Springer International Publishing, 2020, pp. 1–12.

[89] K. Drosou and C. Koukouvinos, "Proximal support vector machine techniques on medical prediction outcome," *Journal of Applied Statistics*, vol. 44, no. 3, pp. 533–553, Feb. 2017, doi: 10.1080/02664763.2016.1177499.

[90] S. M. Kasongo and Y. Sun, "Development and evaluation of a deep learning based intrusion detection model for wireless networks," 2020. https://ujcontent.uj.ac.za/vital/access/manager/Repository/uj:35460?site_name=GlobalView&view=null&f0=sm_identifier%3A%22http%3A%2F%2Fhdl.handle.net%2F10210%2F418298%22&sort=null (accessed Jul. 22, 2020).

[91] "UCI Machine Learning Repository: default of credit card clients Data Set." https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients (accessed Mar. 14, 2020).

[92] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, p. 100203, Jan. 2019, doi: 10.1016/j.imu.2019.100203.

[93] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," *Mobile Information Systems*, vol. 2018, pp. 1–21, Dec. 2018, doi: 10.1155/2018/3860146.

[94] A. M. Alaa, T. Bolton, E. D. Angelantonio, J. H. F. Rudd, and M. van der Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants," *PLOS ONE*, vol. 14, no. 5, p. e0213653, May 2019, doi: 10.1371/journal.pone.0213653.

[95] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 16, Feb. 2020, doi: 10.1186/s12911-020-1023-5.

[96] I. D. Mienye, Y. Sun, and Z. Wang, "Improved sparse autoencoder based artificial neural network approach for prediction of heart disease," *Informatics in Medicine Unlocked*, vol. 18, p. 100307, Jan. 2020, doi: 10.1016/j.imu.2020.100307.

[97] "Framingham Heart study dataset." https://kaggle.com/amanajmera1/framingham-heart-study-dataset (accessed Jan. 24, 2020).

[98] V. M. Valdespino and V. E. Valdespino, "Cervical cancer screening: state of the art," *Curr. Opin. Obstet. Gynecol.*, vol. 18, no. 1, pp. 35–40, Feb. 2006, doi: 10.1097/01.gco.0000192971.59943.89.

[99] K. Rayavarapu and K. K. V. Krishna, "Prediction of Cervical Cancer using Voting and DNN Classifiers," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Coimbatore, Mar. 2018, pp. 1–5, doi: 10.1109/ICCTCT.2018.8551176.

[100] "UCI Machine Learning Repository: Cervical cancer (Risk Factors) Data Set." https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29 (accessed Jan. 27, 2020).

[101] W. Wu and H. Zhou, "Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches," *IEEE Access*, vol. 5, pp. 25189–25195, 2017, doi: 10.1109/ACCESS.2017.2763984.

[102] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer Learning with Partial Observability Applied to Cervical Cancer Screening," in *Pattern Recognition and Image Analysis*, 2017, pp. 243–250.

[103] P. Gupta, I. Jindal, and A. Goyal, "Early Detection and Prevention of Cervical Cancer," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, Mar. 2019, pp. 1–4, doi: 10.1109/I2CT45611.2019.9033800.

[104] A. A. Ogunleye and W. Qing-Guo, "XGBoost Model for Chronic Kidney Disease Diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2019, doi: 10.1109/TCBB.2019.2911071.

[105] L. J. Rubini and P. Eswaran, "UCI Machine Learning Repository: Chronic_Kidney_Disease Data Set," 2015. https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease (accessed Jun. 26, 2020).

[106] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, Aug. 2018, doi: 10.1016/j.aci.2018.08.003.

[107] J. N. Mandrekar, "Receiver Operating Characteristic Curve in Diagnostic Test Assessment," *Journal of Thoracic Oncology*, vol. 5, no. 9, pp. 1315–1316, Sep. 2010, doi: 10.1097/JTO.0b013e3181ec173d.

[108] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," *Caspian J Intern Med*, vol. 4, no. 2, pp. 627–635, 2013, Accessed: Nov. 21, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/.

[109] A. S. Shamsabadi, M. Babaie-Zadeh, S. Z. Seyyedsalehi, H. R. Rabiee, and C. Jutten, "A new algorithm for training sparse autoencoders," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Aug. 2017, pp. 2141–2145, doi: 10.23919/EUSIPCO.2017.8081588.

[110] B. Yan and G. Han, "Effective Feature Extraction via Stacked Sparse Autoencoder to Improve Intrusion Detection System," *IEEE Access*, vol. 6, pp. 41238–41248, 2018, doi: 10.1109/ACCESS.2018.2858277.

[111] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *Proceedings of the 21th international conference on Artificial neural networks - Volume Part I*, Espoo, Finland, Jun. 2011, pp. 44–51, Accessed: Apr. 13, 2020. [Online].

[112] S. Narejo, E. Pasero, and F. Kulsoom, "EEG Based Eye State Classification using Deep Belief Network and Stacked AutoEncoder," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 6, Art. no. 6, Dec. 2016, doi: 10.11591/ijece.v6i6.pp3131-3141.

[113] G. Dong, G. Liao, H. Liu, and G. Kuang, "A Review of the Autoencoder and Its Variants: A Comparative Perspective from Target Recognition in Synthetic-Aperture Radar Images," *IEEE Geoscience and Remote Sensing Magazine*, vol. 6, no. 3, pp. 44–68, Sep. 2018, doi: 10.1109/MGRS.2018.2853555.

[114] "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3784–3797, Aug. 2018, doi: 10.1109/TNNLS.2017.2736643.

[115] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan, "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3637–3647, Oct. 2018, doi: 10.1109/JIOT.2018.2816007.

[116] A. A. Taha and S. J. Malebary, "An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine," *IEEE Access*, vol. 8, pp. 25579–25587, 2020, doi: 10.1109/ACCESS.2020.2971354.

[117] S. Kamley, S. Jaloree, and R. S. Thakur, "Performance Forecasting of Share Market using Machine Learning Techniques: A Review," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 6, Art. no. 6, Dec. 2016, doi: 10.11591/ijece.v6i6.pp3196-3204.

[118] D. Liu, Z. Wang, B. Wen, J. Yang, W. Han, and T. S. Huang, "Robust Single Image Super-Resolution via Deep Networks With Sparse Prior," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3194–3207, Jul. 2016, doi: 10.1109/TIP.2016.2564643.

[119] J. A. Ellis and S. Rajamanickam, "Scalable Inference for Sparse Deep Neural Networks using Kokkos Kernels," in *2019 IEEE High Performance Extreme Computing Conference (HPEC)*, Sep. 2019, pp. 1–7, doi: 10.1109/HPEC.2019.8916378.

[120] H. Asil and J. Bagherzadeh, "A new approach to image classification based on a deep multiclass AdaBoosting ensemble," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 5, Art. no. 5, Oct. 2020, doi: 10.11591/ijece.v10i5.pp4872-4880.
[121] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017, Accessed: Oct. 31, 2019. [Online]. Available: http://arxiv.org/abs/1412.6980.
[122] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv:1502.03167 [cs]*, Mar. 2015, Accessed: Jan. 14, 2020. [Online]. Available: http://arxiv.org/abs/1502.03167.
[123] M. Sun, H. Wang, P. Liu, S. Huang, and P. Fan, "A sparse stacked denoising autoencoder with optimized transfer learning applied to the fault diagnosis of rolling bearings," *Measurement*, vol. 146, pp. 305–314, Nov. 2019, doi: 10.1016/j.measurement.2019.06.029.
[124] H. Zhu, J. Cheng, C. Zhang, J. Wu, and X. Shao, "Stacked pruning sparse denoising autoencoder based intelligent fault diagnosis of rolling bearings," *Applied Soft Computing*, vol. 88, p. 106060, Mar. 2020, doi: 10.1016/j.asoc.2019.106060.
[125] A. Sankaran, M. Vatsa, R. Singh, and A. Majumdar, "Group sparse autoencoder," *Image and Vision Computing*, vol. 60, pp. 64–74, Apr. 2017, doi: 10.1016/j.imavis.2017.01.005.
[126] R. Al-Hmouz, W. Pedrycz, A. Balamash, and A. Morfeq, "Logic-driven autoencoders," *Knowledge-Based Systems*, vol. 183, p. 104874, Nov. 2019, doi: 10.1016/j.knosys.2019.104874.
[127] H. Musafer, A. Abuzneid, M. Faezipour, and A. Mahmood, "An Enhanced Design of Sparse Autoencoder for Latent Features Extraction Based on Trigonometric Simplexes for Network Intrusion Detection Systems," *Electronics*, vol. 9, no. 2, Art. no. 2, Feb. 2020, doi: 10.3390/electronics9020259.
[128] S. Feng, H. Yu, and M. F. Duarte, "Autoencoder based sample selection for self-taught learning," *Knowledge-Based Systems*, vol. 192, p. 105343, Mar. 2020, doi: 10.1016/j.knosys.2019.105343.
[129] B. Xu, H. Lin, Y. Lin, and K. Xu, "Incorporating query constraints for autoencoder enhanced ranking," *Neurocomputing*, vol. 356, pp. 142–150, Sep. 2019, doi: 10.1016/j.neucom.2019.03.068.
[130] O. İrsoy and E. Alpaydın, "Unsupervised feature extraction with autoencoder trees," *Neurocomputing*, vol. 258, pp. 63–73, Oct. 2017, doi: 10.1016/j.neucom.2017.02.075.
[131] J. Xu *et al.*, "Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images," *IEEE Trans Med Imaging*, vol. 35, no. 1, pp. 119–130, Jan. 2016, doi: 10.1109/TMI.2015.2458702.
[132] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, doi: 10.1126/science.1127647.
[133] C. Shi, B. Luo, S. He, K. Li, H. Liu, and B. Li, "Tool Wear Prediction via Multidimensional Stacked Sparse Autoencoders With Feature Fusion," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5150–5159, Aug. 2020, doi: 10.1109/TII.2019.2949355.
[134] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy Layer-Wise Training of Deep Networks," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds. MIT Press, 2007, pp. 153–160.
[135] D. Prusti and S. K. Rath, "Web service based credit card fraud detection by applying machine learning techniques," in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, Oct. 2019, pp. 492–497, doi: 10.1109/TENCON.2019.8929372.
[136] Y. Sayjadah, I. A. T. Hashem, F. Alotaibi, and K. A. Kasmiran, "Credit Card Default Prediction using Machine Learning Techniques," in *2018 Fourth International Conference on Advances in Computing, Communication Automation (ICACCA)*, Oct. 2018, pp. 1–4, doi: 10.1109/ICACCAF.2018.8776802.
[137] T.-C. Hsu, S.-T. Liou, Y.-P. Wang, Y.-S. Huang, and Che-Lin, "Enhanced Recurrent Neural Network for Combining Static and Dynamic Features for Credit Card Default Prediction," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 1572–1576, doi: 10.1109/ICASSP.2019.8682212.

[138] W. A. Chishti and S. M. Awan, “Deep Neural Network a Step by Step Approach to Classify Credit Card Default Customer,” in *2019 International Conference on Innovative Computing (ICIC)*, Nov. 2019, pp. 1–8, doi: 10.1109/ICIC48496.2019.8966723.
[139] D. E. Stanley and D. G. Campos, “The Logic of Medical Diagnosis,” *Perspectives in Biology and Medicine*, vol. 56, no. 2, pp. 300–315, Aug. 2013, doi: 10.1353/pbm.2013.0019.
[140] H. M. Epstein, “The most important medical issue ever: And why you need to know more about it,” *Society to Improve Diagnosis in Medicine*, 2019. https://www.improvediagnosis.org/dxiq-column/most-important-medical-issue-ever/ (accessed Aug. 30, 2020).
[141] N. Liu, X. Li, E. Qi, M. Xu, L. Li, and B. Gao, “A novel Ensemble Learning Paradigm for Medical Diagnosis with Imbalanced Data,” *IEEE Access*, pp. 1–1, 2020, doi: 10.1109/ACCESS.2020.3014362.
[142] Z. Ma *et al.*, “Lightweight Privacy-preserving Medical Diagnosis in Edge Computing,” *IEEE Transactions on Services Computing*, pp. 1–1, 2020, doi: 10.1109/TSC.2020.3004627.
[143] X. Li, M. Jia, M. T. Islam, L. Yu, and L. Xing, “Self-supervised Feature Learning via Exploiting Multi-modal Data for Retinal Disease Diagnosis,” *IEEE Transactions on Medical Imaging*, pp. 1–1, 2020, doi: 10.1109/TMI.2020.3008871.
[144] Z. Chen, R. Guo, Z. Lin, T. Peng, and X. Peng, “A data-driven health monitoring method using multi-objective optimization and stacked autoencoder based health indicator,” *IEEE Transactions on Industrial Informatics*, pp. 1–1, 2020, doi: 10.1109/TII.2020.2999323.
[145] U. Raghavendra, A. Gudigar, S. V. Bhandary, T. N. Rao, E. J. Ciaccio, and U. R. Acharya, “A Two Layer Sparse Autoencoder for Glaucoma Identification with Fundus Images,” *J Med Syst*, vol. 43, no. 9, p. 299, Jul. 2019, doi: 10.1007/s10916-019-1427-x.
[146] L. Verma, S. Srivastava, and P. C. Negi, “A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data,” *J Med Syst*, vol. 40, no. 7, p. 178, Jun. 2016, doi: 10.1007/s10916-016-0536-z.
[147] B. A. Tama, S. Im, and S. Lee, “Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble,” *BioMed Research International*, Apr. 27, 2020. https://www.hindawi.com/journals/bmri/2020/9816142/ (accessed Aug. 28, 2020).
[148] E. Ahishakiye, R. Wario, W. Mwangi, and D. Taremwa, “Prediction of Cervical Cancer Basing on Risk Factors using Ensemble Learning,” in *2020 IST-Africa Conference (IST-Africa)*, May 2020, pp. 1–12.
[149] Y. Xiong and Y. Lu, “Deep Feature Extraction From the Vocal Vectors Using Sparse Autoencoders for Parkinson’s Classification,” *IEEE Access*, vol. 8, pp. 27821–27830, 2020, doi: 10.1109/ACCESS.2020.2968177.
[150] M. Daoud, M. Mayo, and S. J. Cunningham, “RBFA: Radial Basis Function Autoencoders,” in *2019 IEEE Congress on Evolutionary Computation (CEC)*, Jun. 2019, pp. 2966–2973, doi: 10.1109/CEC.2019.8790041.
[151] A. Ng, “Sparse autoencoder.” 2011, Accessed: Jun. 06, 2020. [Online]. Available: https://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf.
[152] K. Kayabol, “Approximate Sparse Multinomial Logistic Regression for Classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 490–493, Feb. 2020, doi: 10.1109/TPAMI.2019.2904062.
[153] J. L. L. Herrera, H. V. R. Figueroa, and E. J. R. Ramírez, “Deep fraud. A fraud intention recognition framework in public transport context using a deep-learning approach,” in *2018 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, Feb. 2018, pp. 118–125, doi: 10.1109/CONIELECOMP.2018.8327186.
[154] S. Ruder, “An overview of gradient descent optimization algorithms,” *arXiv:1609.04747 [cs]*, Jun. 2017, Accessed: Oct. 30, 2020. [Online]. Available: http://arxiv.org/abs/1609.04747.
[155] N. Qian, “On the momentum term in gradient descent learning algorithms,” *Neural Networks*, vol. 12, no. 1, pp. 145–151, Jan. 1999, doi: 10.1016/S0893-6080(98)00116-6.

[156] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telematics and Informatics*, vol. 36, pp. 82–93, Mar. 2019, doi: 10.1016/j.tele.2018.11.007.
[157] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
[158] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, Apr. 2019, pp. 292–297, doi: 10.1109/ICOEI.2019.8862604.
[159] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," *Expert Systems with Applications*, vol. 68, pp. 163–172, Feb. 2017, doi: 10.1016/j.eswa.2016.10.020.
[160] Abdullah, F. B. Ashraf, and N. S. Momo, "Comparative analysis on Prediction Models with various Data Preprocessings in the Prognosis of Cervical Cancer," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2019, pp. 1–6, doi: 10.1109/ICCCNT45670.2019.8944850.
[161] C.-C. Chang, S.-L. Cheng, C.-J. Lu, and K.-H. Liao, "Prediction of Recurrence in Patients with Cervical Cancer Using MARS and Classification," *International Journal of Machine Learning and Computing*, pp. 75–78, 2013, doi: 10.7763/IJMLC.2013.V3.276.
[162] M. F. Ijaz, M. Attique, and Y. Son, "Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods," *Sensors*, vol. 20, no. 10, Art. no. 10, Jan. 2020, doi: 10.3390/s20102809.
[163] B. Nithya and V. Ilango, "Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction," *SN Appl. Sci.*, vol. 1, no. 6, p. 641, May 2019, doi: 10.1007/s42452-019-0645-7.
[164] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics in Medicine Unlocked*, vol. 15, p. 100178, Jan. 2019, doi: 10.1016/j.imu.2019.100178.
[165] D. Gupta, S. Khare, and A. Aggarwal, "A method to predict diagnostic codes for chronic diseases using machine learning techniques," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, Apr. 2016, pp. 281–287, doi: 10.1109/CCAA.2016.7813730.
[166] B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, "An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy," *IEEE Access*, vol. 8, pp. 55012–55022, 2020, doi: 10.1109/ACCESS.2020.2981689.
[167] N. V. G. Raju, K. P. Lakshmi, K. G. Praharshitha, and C. Likhitha, "Prediction of chronic kidney disease (CKD) using Data Science," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, May 2019, pp. 642–647, doi: 10.1109/ICCS45141.2019.9065309.
[168] A. J. Aljaaf *et al.*, "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics," in *2018 IEEE Congress on Evolutionary Computation (CEC)*, Jul. 2018, pp. 1–9, doi: 10.1109/CEC.2018.8477876.

UNIVERSITY OF JOHANNESBURG